Elucidation du métabolisme des microorganismes par la modélisation et l'interprétation des données d'essentialité de gènes. Application au métabolisme de la bactérie Acinetobacter baylyi ADP1.

Maxime Durot

2009EVRY0017

ÉLUCIDATION DU METABOLISME DES MICROORGANISMES PAR LA MODELISATION ET L'INTERPRETATION DES DONNEES D'ESSENTIALITE DE GENES.

APPLICATION AU METABOLISME DE LA BACTERIE *ACINETOBACTER BAYLYI* ADP1.

MAXIME DUROT

Thèse de Doctorat

Spécialité : Bioinformatique, biologie structurale et génomique

Université Evry Val d'Essonne

École doctorale : Des génomes aux organismes

Soutenue le 12 octobre 2009 devant le jury composé de :

Jean-Pierre MAZAT	rapporteur
Stefan SCHUSTER	rapporteur
Antoine DANCHIN	examinateur
Eytan RUPPIN	examinateur
Vincent SCHACHTER	directeur de thèse
Jean WEISSENBACH	directeur de thèse

RESUME

Deux échelles d'observations sont traditionnellement utilisées pour étudier le métabolisme des microorganismes: d'une part, à l'échelle locale, la caractérisation individuelle des réactions ayant lieu dans la cellule et d'autre part, à l'échelle globale, l'étude de la physiologie de la cellule. Ces deux échelles ont bénéficié de progrès technologiques récents : l'analyse des génomes séquencés permet d'identifier une large fraction des enzymes catalysant les réactions ; la physiologie des microorganismes peut être étudiée à haut débit pour de nombreux environnements et perturbations génétiques. Cependant, l'exploitation conjointe de ces deux échelles demeure complexe car le comportement physiologique global de la cellule résulte de l'action coordonnée de nombreuses réactions. Les approches de modélisation mathématique ont toutefois récemment permis de relier ces deux échelles à l'aide de modèles globaux du métabolisme.

Dans cette thèse, nous explorerons l'utilisation de ces modèles pour compléter la connaissance des réactions à l'aide d'une catégorie particulière de données d'échelle globale : les essentialités de gènes déterminées en observant les phénotypes de croissance de mutants de délétion. Nous nous appuierons pour cela sur la bactérie *Acinetobacter baylyi* ADP1 pour laquelle une collection complète de mutants de délétion a été récemment constituée au Genoscope.

Après avoir présenté les étapes clés et les développements que nous avons effectués pour reconstruire un modèle global du métabolisme d'*A. baylyi*, nous montrerons que la confrontation entre phénotypes observés et phénotypes prédits permet de mettre en évidence des incohérences entre les deux échelles d'observations. Nous montrerons ensuite qu'une interprétation formelle de ces incohérences permet de corriger le modèle et d'améliorer la connaissance du métabolisme. Nous illustrerons ce propos en présentant les corrections que nous avons réalisées à l'aide des phénotypes de mutants d'*A. baylyi*. Enfin, dans une dernière partie, nous proposerons une méthode permettant d'automatiser la correction des incohérences causées par des erreurs d'association entre gènes et réactions.

ABSTRACT

Model-based investigation of microbial metabolism to interpret gene essentiality results, illustrated on *Acinetobacter baylyi* ADP1 metabolism.

Microbial metabolism has traditionally been investigated at two different scales: the finest involves characterizing individually each reaction occurring in the cell; the largest focuses on global cell physiology. Both scales have recently benefited from technological advances: analyzing sequenced genomes identifies a large fraction of reaction-catalyzing enzymes; cell physiology can be determined at high-throughput for several environmental conditions and genetic perturbations. Combining both scales remains, however, especially complex as the global physiological behavior of a cell results from the coordinated action of a large network of reactions. Mathematical modeling approaches have yet shown recently that genome-scale metabolic models could help in linking both scales.

In this thesis, we explore the use of such models to expand the knowledge of reactions with a specific type of high-level data: gene essentiality data, assessed using growth phenotypes of deletion mutants. We will use as model organism the bacterium *Acinetobacter baylyi* ADP1, for which a genome-wide collection of gene deletion mutants has recently been created.

Following a presentation of the key steps and developments that have been required to reconstruct a global metabolic model of *A. baylyi*, we will show that confronting observed and predicted phenotypes highlight inconsistencies between the two scales. We will then show that a formal interpretation of these inconsistencies can guide model corrections and improvements to the knowledge of metabolism. We will illustrate this claim by presenting model corrections triggered by *A. baylyi* mutant phenotypes. Finally, we will introduce a method that automates the correction of inconsistencies caused by wrong associations between genes and reactions.

REMERCIEMENTS

Je tiens à remercier en premier lieu Vincent Schachter, pour m'avoir tout d'abord convaincu d'entreprendre cette thèse puis guidé scientifiquement ces quatre années. Il aura été le garant de la présence de développements méthodologiques et théoriques dans mes travaux, sachant me faire prendre du recul à bon escient lorsqu'il m'arrivait de me perdre dans les détails de la biochimie d'*Acinetobacter baylyi*. Professionnellement, je lui suis largement redevable de m'avoir introduit dans la vie scientifique internationale à travers les collaborations, projets européens, séminaires et conférences auxquels il m'a associé.

Je remercie de même Jean Weissenbach pour avoir accepté de diriger ma thèse et permis le développement de mon sujet de recherche, relativement original au Genoscope. Mes travaux se sont fondés sur les nombreux échanges qu'il aura su favoriser avec les équipes expérimentales du laboratoire.

Un très grand merci à tous les membres de l'équipe Nemo, présents et passés, avec qui j'ai travaillé au quotidien et pu échanger des idées sur mes travaux : F. Le Fèvre, B. Pinaud, S. Smidtas, C. Combe, M. Heinig, V. Sabarly, P-Y. Bourguignon, G. Vieira et R. Baran. Merci en particulier à François Le Fèvre avec qui j'ai partagé la lourde tâche de parcourir le métabolisme entier d'*A. baylyi* et pour ses encouragements de collègue de bureau.

Je remercie vivement l'ensemble de l'équipe Thesaurus Métabolique du Genoscope, et en particulier Véronique de Berardinis et Marcel Salanoubat, pour

avoir apporté de la « réalité expérimentale » à mes travaux. Merci d'avoir passé de longues heures à m'aider à mieux comprendre les habitudes d'*A. baylyi* et de ses mutants !

Je remercie également Alain Perret et Christophe Lechaplais pour leurs contributions expérimentales à cette thèse, ainsi qu'Annett Kreimeyer et Georges Cohen pour avoir pris le temps de puiser dans leur formidable connaissance du métabolisme pour répondre à mes questions.

Merci à l'Atelier de Génomique Comparative, et en particulier à David Vallenet pour m'avoir donné une *loupe* pour explorer les génomes bactériens et à Claudine Médigue pour m'avoir permis de conclure ma thèse dans son équipe.

L'aide de l'équipe informatique du Genoscope m'aura souvent été précieuse, merci à eux pour leur support et leurs conseils.

Je remercie les membres du jury pour m'avoir fait l'honneur de leur présence à ma soutenance et m'avoir aidé, par leur remarques et conseils, à améliorer mon manuscrit.

Je suis très reconnaissant envers le Genoscope et le CEA pour m'avoir permis de réaliser cette thèse conjointement avec mes activités professionnelles.

Enfin, un grand merci pour leur soutien sans faille à mes parents, ma sœur, ma belle-famille et l'ensemble de mes proches que je ne saurai lister ici. Et, plus que tout, merci à ma femme, Marie-Perrine, pour son amour qui aura toujours su me remotiver dans les moments difficiles et pour avoir mené de front avec succès préparation de mariage et soutien de conjoint en rédaction de thèse !

TABLE DES MATIERES

AVANT-PROPOS

Les organismes vivants sont tous de formidables chimistes aux capacités souvent insoupçonnées. Chaque cellule est le siège d'un nombre considérable de réactions qui lui permettent de créer les molécules nécessaires à sa vie à partir des molécules de son environnement. Cet ensemble de réactions biochimiques, que l'on appelle le *métabolisme* des cellules, a attiré depuis longtemps la curiosité de l'homme. Non seulement, d'un point de vue fondamental, il est essentiel d'aborder la chimie des cellules pour en comprendre leur fonctionnement et leurs interactions avec le milieu extérieur, mais également, d'un point de vue pratique, l'utilisation de leurs métabolismes occupe une place significative dans les activités humaines. De la fermentation alcoolique à la synthèse de biocarburants en passant par l'épuration des eaux usées, les compétences biochimiques des organismes offrent des solutions technologiques à de nombreux besoins.

Cette thèse aborde l'étude du métabolisme de manière pluridisciplinaire, associant biochimie, génétique et modélisation mathématique. Traditionnellement, deux échelles d'observations sont utilisées pour appréhender le métabolisme. D'une part, les approches classiques de biochimie permettent de caractériser la chimie des réactions ayant lieu dans les cellules. Ainsi au cours des dernières décennies et encore aujourd'hui, un nombre croissant de réactions métaboliques sont élucidées de cette manière, principalement chez les quelques organismes modèles. D'autre part, à une échelle plus grande, l'observation de la physiologie des cellules permet d'en caractériser la biochimie de manière globale : par exemple quelles molécules

extérieures sont requises et en quelles proportions pour permettre la croissance. Bien que présentant le métabolisme sous deux échelles différentes, associer ces deux types d'observations n'est pas chose simple. Le grand nombre et la complexité des enchaînements de réactions métaboliques rendent en effet difficile la déduction de caractéristiques métaboliques globales à partir de la seule connaissance des réactions le composant. Dans ce but, des modèles mathématiques du métabolisme ont récemment été introduits pour effectuer ce raisonnement de manière appropriée. Cette thèse se propose d'approfondir l'utilisation des modèles du métabolisme dans l'objectif d'élucider au mieux le métabolisme de microorganismes encore peu étudiés en exploitant conjointement données physiologiques globales et caractérisations locales de réactions.

Ce type d'approche est aujourd'hui rendu possible grâce à des avancées technologiques récentes. D'une part, alors que les techniques expérimentales traditionnelles de biochimie ont un débit beaucoup trop faible pour détecter exhaustivement les réactions métaboliques de nouveaux organismes, le séquençage et l'annotation de leurs génomes offrent une solution alternative efficace. L'avènement des méthodes comparatives permet en effet de déduire la fonction biochimique d'une proportion significative des gènes par homologie aux gènes connus chez les autres organismes, et d'inférer ainsi une grande partie de ses réactions métaboliques. Mais l'utilisation exclusive de ces méthodes trouve rapidement ses limites pour des activités biochimiques spécifiques à l'organisme ou encore peu étudiées. D'autre part, le débit des expériences sur la physiologie des organismes a également augmenté récemment, en particulier pour les microorganismes. Nous utiliserons une catégorie particulière de ces expériences, mêlant à grande échelle perturbation génétique et caractérisation physiologique. Elles consistent à créer systématiquement un mutant de délétion pour chacun des gènes d'un organisme. La capacité ou non de croître de chacun de ces mutants dans des environnements chimiques donnés (leurs *phénotypes* de croissance) offre une information utile quant au rôle du gène délété – et par extension de la fonction biochimique inactivée – dans le métabolisme de la bactérie. Cette thèse explore spécifiquement l'utilisation des modèles du métabolisme pour compléter la connaissance du métabolisme obtenue par les données de séquences avec les phénotypes de croissance expérimentaux. La bactérie *Acinetobacter baylyi* ADP1

(que nous nommerons simplement *A. baylyi*) nous accompagnera tout au long de ce manuscrit, se prêtant comme sujet d'étude à la fois *in vivo* et *in silico*.

La première partie de ce manuscrit introduit les notions manipulées dans la thèse : le métabolisme, les expériences de génétique et la modélisation mathématique du métabolisme. Cette partie cherche à balayer l'état de l'art dans ces trois domaines et à placer la contribution de la thèse dans le contexte des travaux antérieurs pertinents.

Dans une deuxième partie, nous présenterons de manière détaillée la reconstruction du modèle métabolique global d'*A. baylyi* à partir de son annotation génomique. Cette section décrit naturellement le processus ayant permis d'identifier les activités métaboliques présentes chez cette bactérie, mais également les spécificités associées à la modélisation retenue. Il nous a semblé en effet important de nous attarder sur les hypothèses de modélisation et leurs conséquences sur la construction des modèles. Alors même que de nombreux articles de revue présentent comment reconstruire des voies métaboliques à partir d'une annotation d'un génome, peu d'entre eux détaillent les points clés liés à la modélisation.

La troisième partie du manuscrit aborde l'exploitation des phénotypes de mutants par les modèles métaboliques. Nous montrerons, toujours sur la base du métabolisme d'*A. baylyi*, qu'en identifiant les incohérences entre les phénotypes prédits par le modèle et les phénotypes observés, des erreurs dans la connaissance du métabolisme peuvent être pointées précisément. Nous verrons dans quelle mesure ces erreurs peuvent être corrigées à l'aide de ces données. Nous discuterons également à cette occasion de la notion d'*essentialité* des gènes, et de ses liens avec le métabolisme et l'environnement de la cellule.

La quatrième partie traite de l'automatisation de l'interprétation de ces incohérences lorsqu'elles sont d'origine génétique. À travers une formalisation rigoureuse du raisonnement portant sur l'association entre gènes et réactions, nous montrerons qu'il est possible de déduire automatiquement les associations gènes - réactions qui soient compatibles avec les phénotypes de mutants observés. Ces raisonnements retrouvent une partie des interprétations effectuées « manuellement » et forment une brique indispensable à l'interprétation métabolique à grande échelle des phénotypes de mutants.

Enfin, dans une dernière partie, nous reprendrons les principales conclusions de nos travaux et les mettrons en perspective des évolutions de la discipline. La thématique de la thèse étant en plein essor, nous réeffectuerons un tour d'horizon des travaux similaires publiés à la fin de la thèse. Plus largement, nous discuterons également de la place d'approches de modélisation dans la reconstruction du métabolisme de nouveaux organismes, à l'heure où le débit des nouvelles technologies permet de séquencer un génome bactérien en quelques jours.

INTRODUCTION

Ce chapitre a pour but d'introduire au lecteur les concepts biologiques et mathématiques utilisés dans cette thèse et d'effectuer un état de l'art dans les domaines couverts. Nous l'avons divisé en cinq parties. La première s'attache à introduire les notions utiles à la compréhension du métabolisme des microorganismes ainsi qu'à présenter l'état de l'art quant à son exploration. La deuxième partie se concentrera sur l'utilisation des phénotypes de croissance pour étudier le métabolisme et en particulier aux techniques de génétique à haut débit associées. Dans la troisième partie, le lecteur trouvera une revue actuelle des méthodes de modélisation mathématique appliquées au métabolisme, ainsi qu'une présentation détaillée du cadre de modélisation que nous avons retenu : la modélisation à base de contrainte. Dans la quatrième, nous présenterons les caractéristiques et les ressources disponibles sur l'organisme modèle utilisé dans cette thèse, *Acinetobacter baylyi* ADP1. Enfin, nous effectuerons en dernier lieu une synthèse de l'état de l'art et présenterons le sujet de notre thèse dans ce contexte.

1 Le métabolisme : la chimie du vivant

1.1 Quelques faits remarquables

Une des caractéristiques majeures des organismes vivants est leur aptitude à croître et à se reproduire par eux-mêmes. Pour ce faire, les processus mis en œuvre sont en grande majorité de nature chimique (*bio*chimique), impliquant une grande variété de molécules. On désigne généralement par *métabolisme* les processus

biochimiques ayant pour rôle la synthèse et la dégradation de ces biomolécules ainsi que la transformation d'énergie chimique. Cette définition distingue ainsi le métabolisme d'autres processus chimiques à l'œuvre dans les cellules, tels que la signalisation, la réplication et la transcription de l'ADN, ou l'assemblage des protéines.

Le métabolisme est indispensable à la vie. D'un point de vue thermodynamique, les organismes vivants sont des systèmes fondamentalement hors d'équilibre qui nécessitent pour maintenir cet état d'échanger continuellement de l'énergie et de la matière avec le milieu extérieur (nous aborderons ce point plus en détails section 1.2.6). Le métabolisme joue un rôle essentiel dans cet échange d'énergie et de matière. Cependant, toutes les entités vivantes ne possèdent pas nécessairement de métabolisme propre, encore que les nombreuses définitions d' « être vivant » soient parfois associées à sa présence[1]. C'est le cas des virus et dans une moindre mesure de certaines bactéries parasites ; ceux-ci exploitent directement les ressources de leurs hôtes. À titre d'exemple, la bactérie parasite *Rickettsia prowazekii*, qui vit majoritairement dans le cytoplasme de son hôte, dépend très fortement du métabolisme de ce dernier ; elle ne peut synthétiser elle-même la plupart de ses constituants et profite dès que possible de l'énergie chimique de son hôte (Andersson et al. 1998).

Néanmoins, dans leur très grande majorité, les cellules des organismes vivants consacrent une grande partie de leurs activités à exploiter et à transformer les molécules de leur entourage (leur *environnement*) pour en retirer de l'énergie et créer les molécules qui serviront à leur propre construction. Ce sont ces réactions qui font des organismes vivants de véritables chimistes.

[1] Voir par exemple les nombreuses définitions proposées dans l'article Wikipedia sur les organismes vivants : http://en.wikipedia.org/wiki/Life#Definitions .

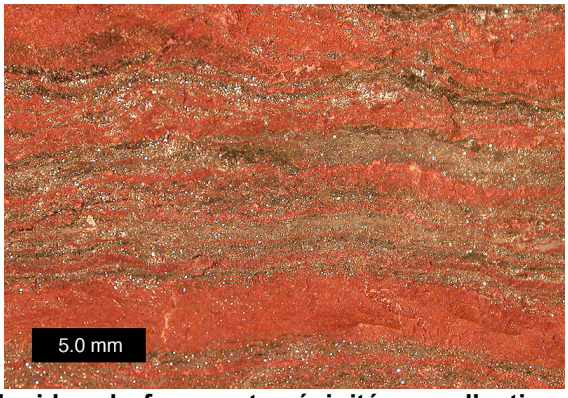


Figure 1. Couches d'oxides de fer ayant précipité sous l'action de l'oxygène produit par la photosynthèse. Photographie d'un échantillon issu de la péninsule supérieure du Michigan (source http://en.wikipedia.org/wiki/Banded_iron_formation)

Le volume d'action du métabolisme peut être considérable. Pour ne prendre qu'un exemple, nous rappellerons au lecteur qu'une très grande majorité du dioxygène présent dans l'atmosphère terrestre est d'origine « biologique». L'apparition de la photosynthèse dans l'arsenal métabolique du vivant a en effet modifié significativement la composition de l'atmosphère, il y a environ deux milliards d'années (Knoll 2003). La production massive de dioxygène par les organismes photosynthétiques transforma alors l'atmosphère réductrice en une atmosphère oxydante, laissant des traces visibles dans les couches géologiques de l'époque (voir Figure 1). On estime que le flux actuel de création de dioxygène par la photosynthèse permettrait de régénérer l'ensemble de l'oxygène atmosphérique en 2000 ans (Dole 1965).

Le métabolisme marque également par sa diversité. Certains organismes, et en particulier des bactéries, ont été découverts dans des environnements très variés, au sein desquels les molécules sources d'énergie et de matière diffèrent de manière considérable. À titre illustratif, pour générer leur énergie, les bactéries tirent parti de diverses manières des potentiels d'oxydoréduction des molécules de leur environnement. Tandis que dans les milieux aérobies courants, les molécules organiques sont généralement oxydées en utilisant l'oxygène comme accepteur d'électron, en milieu anaérobie certains organismes remplacent ce dernier par d'autres molécules organiques (par exemple lors de la fermentation) ou des formes oxydées de l'azote (ex. : nitrate, nitrite), du soufre (ex. : sulfate ou sulfite) ou de métaux (ex. : fer, manganèse, voire même certains métaux lourds). À l'inverse, on a découvert des

organismes pouvant remplacer les molécules organiques par d'autres donneurs d'électrons[2]. Ces organismes génèrent leur énergie en oxydant par exemple les molécules réduites de dihydrogène, de soufre (inorganique par ex.), d'azote (ammoniaque) ou de fer.

Le répertoire des molécules organiques pouvant être « métabolisées » est lui-même extrêmement large. On estime qu'environ un millier de molécules composent le métabolisme primaire[3] de la majorité des organismes. À cet ensemble, les organismes supérieurs – en particulier les plantes et les champignons – ajoutent les molécules de leur métabolisme secondaire[4] dont on estime la diversité à plusieurs centaines de milliers (Villas-Boas et al. 2007, pp.25-26). Les structures de ces molécules sont souvent remarquablement complexes (voir Figure 2), leurs rôles biologiques dépendant en grande partie de ces structures et se révélant parfois extrêmement sensibles à tout changement de chiralité[5]. À cet effet, certaines voies de synthèse du métabolisme sont particulièrement efficaces à produire spécifiquement certains énantiomères donnés.

[2] On les nomme *lithotrophes*, par opposition aux *organotrophes*.

[3] Le métabolisme primaire regroupe les activités métaboliques participant au développement et à la croissance de l'organisme, telles que la génération d'énergie et la synthèse des constituants de la cellule. Ces activités sont relativement ubiquitaires entre les organismes.

[4] Le métabolisme secondaire regroupe les activités de synthèse de molécules ne contribuant pas directement à la croissance de la cellule. Ces molécules ont par exemple des rôles dans la communication ou les interactions écologiques.

[5] Une molécule est *chirale* si elle n'est pas superposable à son image dans un miroir. Les deux molécules images l'une de l'autre sont alors appelées *énantiomères*. Deux énantiomères ont des formules développées identiques mais ont des structures tridimensionnelles distinctes. Cette différence peut leur conférer des propriétés physiques, chimiques ou biologiques distinctes.

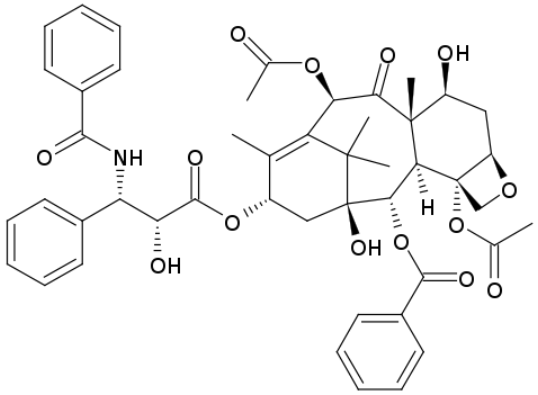


Figure 2. La molécule de Taxol, utilisée en chimiothérapie. Cette molécule a été découverte dans l'écorce d'une espèce d'if, *Taxus brevifolia*.

Similairement à ces capacités de synthèse, les organismes ont développé un ensemble de réactions leur permettant de dégrader et d'utiliser à leur avantage un large spectre de molécules. Ceci est notamment vrai pour les bactéries, lesquelles ont développé un ensemble de stratégies pour croître dans des environnements chimiques variés voire extrêmes. Leurs remarquables capacités d'adaptation les ont même amenées à exploiter des molécules non naturelles produites par l'homme (molécules xénobiotiques), tels que des composés organochlorés ou polyaromatiques (Janssen et al. 2005; van der Meer et al. 1992) .

L'homme utilise depuis longtemps les compétences biochimiques des organismes. Depuis leur utilisation pour la production de fromage, de bière et de vin par fermentation (dont on retrouve des traces de pratique datant de la préhistoire (McGovern et al. 1996)), les applications du métabolisme des microorganismes se sont étendues à de nombreux autres domaines. La pratique de l'ingénierie du métabolisme permet de produire efficacement une large gamme de produits par voie biologique : compléments alimentaires, substances énergétiques, solvants, antibiotiques, vitamines, polymères, pigments (Stephanopoulos et al. 1998, pp.203-283). La voie de production biologique prend surtout son sens lorsque la synthèse chimique se révèle difficile et coûteuse, comme cela est le cas par exemple pour le 1,3-propanediol (Tong et al. 1991), un précurseur de nombreux polymères à forte valeur ajoutée, ou l'artémisinine (Ro et al. 2006), une molécule active contre le paludisme. Les capacités de dégradation des microorganismes sont également utilisées à des fins pratiques, l'exemple le plus flagrant étant leur utilisation primordiale dans les processus d'épuration des eaux usées. L'aptitude des

microorganismes à s'adapter pour utiliser des substances variées en fait des candidats prometteurs pour dégrader des polluants complexes, tels que les polychlorobiphényles (PCB) ou les mélanges de benzène, toluène et xylène (BTX) (Stephanopoulos et al. 1998, pp.266-273).

1.2 Les acteurs du métabolisme

Avant de présenter plus en détail l'organisation du métabolisme au sein des organismes, nous allons préalablement définir dans cette section les « acteurs » impliqués. Nous rappellerons en outre au lecteur quelques notions physiques en rapport avec les réactions biochimiques. En effet, le comportement du métabolisme découle *in fine* de ces notions physiques ; les modèles mathématiques du métabolisme s'appuient de ce fait de manière fondamentale sur la physique à l'œuvre, aux échelles à la fois de la molécule (description des réactions) et de la cellule (cinétique et thermodynamique).

1.2.1 Métabolites

On utilise généralement le terme de *métabolite* pour désigner les molécules impliquées dans le métabolisme cellulaire. Ces molécules sont, dans leur grande majorité, des molécules organiques, composées de carbone et d'hydrogène mais également d'oxygène et dans une moindre mesure d'azote, de phosphore et de soufre. À titre illustratif, la composition moyenne de la bactérie *Lactobacillus lactis* en ces éléments (relativement au carbone) a été évaluée à $C_1H_{1,9}O_{0,6}N_{0,2}P_{0,02}S_{0,01}$ (Oliveira et al. 2005). Cette composition n'est pas fixe et évolue notamment en fonction de l'environnement de croissance de l'organisme, mais elle est indicative de l'ordre de grandeur de la répartition de ces éléments[6]. La forte proportion du carbone dans la composition des métabolites n'est pas anodine. En effet, les propriétés électroniques

[6] On retrouve en réalité d'autres éléments dans la composition des cellules, souvent en moindre quantité. Ce sont principalement des ions jouant le rôle d'électrolytes afin de maintenir une pression osmotique et un pH constants et de favoriser l'import de métabolites (potassium, sodium, calcium, manganèse, chlore). De nombreux métaux de transition (fer, zinc, manganèse, molybdène, cuivre, cobalt, nickel) sont également présents à l'état de trace ; ils sont néanmoins essentiels à l'activité de certaines enzymes.
Cependant, dans la très grande majorité des cas, ces éléments n'entrent pas dans la composition des métabolites.

de l'atome de carbone font qu'il établit facilement jusqu'à quatre liaisons covalentes relativement solides ; cette caractéristique lui permet de générer une combinatoire extrêmement grande de molécules organiques en assemblant plusieurs atomes de carbone entre eux.

1.2.2 Réactions

Les métabolites se transforment chimiquement au cours des *réactions métaboliques* : des métabolites *substrats* réagissent entre eux pour donner des métabolites *produits*. On représente généralement la réaction par son *équation bilan*, laquelle met en évidence la *stœchiométrie* de la réaction, c'est-à-dire les proportions dans lesquelles les métabolites sont consommés et produits (voir Figure 3). L'équation bilan répertorie exhaustivement les substrats et produits impliqués par la réaction. De ce fait, et étant donné que les transformations à l'œuvre sont purement chimiques – celles-ci mettent uniquement en jeu des échanges d'atomes ou de groupes d'atomes entre métabolites par modification de leurs liaisons chimiques – la quantité de chaque élément et la charge globale est conservée : l'équation bilan est dite *équilibrée*.

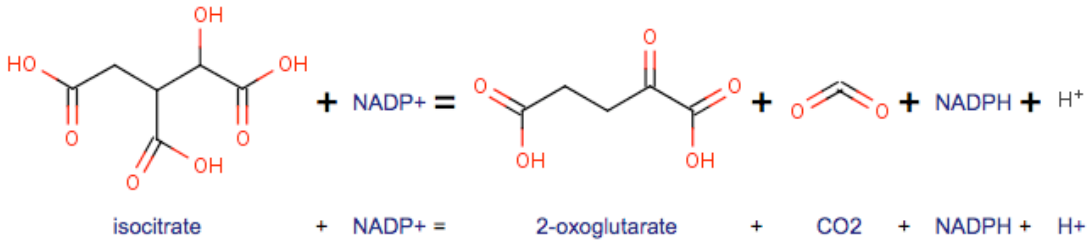


Figure 3. Equation bilan de la réaction catalysée par l'enzyme isocitrate dehydrogénase. Extrait de BRENDA (http://www.brenda-enzymes.info).

On distingue souvent deux types de métabolites dans une réaction biochimique : les substrats et produits *principaux* d'une part et les *cofacteurs* (ou *coenzymes*) d'autre part. Le premier type désigne les métabolites directement transformés par la réaction chimique : il s'agit par exemple de l'isocitrate, du 2-oxoglutarate et du CO_2 dans le cas de la réaction présentée sur la Figure 3. Les cofacteurs désignent quant à eux les métabolites aidant la transformation chimique principale, que ce soit en apportant de l'énergie, en agissant comme accepteur ou donneur d'électron (tels que $NADP^+$ et NADPH dans la réaction de la Figure 3) ou en favorisant le transfert de groupements chimiques. Les transformations chimiques des cofacteurs sont réversibles et, comme

nous le verrons plus loin, une partie des activités métaboliques de la cellule consiste justement à régénérer les cofacteurs en les retransformant dans leur état initial.

1.2.3 Enzymes

Aux cotés des métabolites, les *enzymes* constituent le deuxième acteur clé du métabolisme. Ces dernières jouent en effet le rôle de catalyseurs sans lesquels la plupart des réactions métaboliques ne pourraient se dérouler à des vitesses compatibles avec la vie de la cellule. Le principe de la catalyse enzymatique repose sur une interaction entre l'enzyme et les substrats qui favorise la stabilisation de l'état de transition de la réaction (Koshland 1958). Cette stabilisation abaisse l'énergie à fournir pour atteindre l'état de transition (*énergie d'activation*) et, de ce fait, un nombre plus élevé de substrats d'énergie moindre pourront interagir, accélérant ainsi la réaction (voir Figure 4).

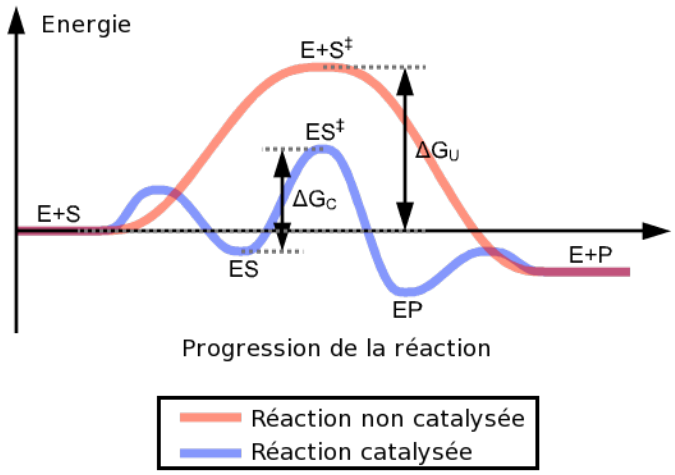


Figure 4. Illustration de la diminution d'énergie d'activation d'une réaction par catalyse enzymatique. E, enzyme ; S, substrat ; $S^{‡}$, état de transition ; P, produit ; ΔG, énergie d'activation avec (ΔG_C) ou sans (ΔG_U) catalyse. Adapté de Wikipedia[7].

Des mécanismes enzymatiques relativement différents permettent d'abaisser l'énergie d'activation, allant d'une stabilisation par effet électrostatique au rapprochement forcé des substrats. Nous n'entrerons cependant pas dans leurs détails qui seraient hors du propos de cette thèse.

Il est cependant important de noter que, dans la grande majorité des cas, les enzymes catalysent des réactions spécifiques alors que les métabolites peuvent

[7] http://en.wikipedia.org/wiki/Enzyme_catalysis

généralement réagir entre eux de diverses manières. En abaissant l'énergie d'activation pour un chemin réactionnel donné – en stabilisant par exemple un état de transition particulier – les enzymes favorisent alors spécifiquement une réaction particulière par rapport aux autres. Le caractère *spécifique* de la catalyse enzymatique est au moins aussi important dans le métabolisme que l'accélération de la vitesse des réactions. Il lui permet en effet d'assurer la transformation des métabolites en des produits particuliers, évitant la production d'autres produits qui en réduiraient le rendement et pourraient s'avérer néfastes. En résumé, le double aspect *spécificité* et *accélération* de la catalyse enzymatique donne à l'organisme le contrôle des transformations métaboliques se déroulant dans la cellule.

À la grande variété de réactions métaboliques correspond une grande variété d'enzymes. Afin d'organiser la description des enzymes identifiées, l'*International Union of Biochemistry and Molecular Biology* (IUBMB)[8] élabore une classification des enzymes basée sur le type de réaction catalysée : la *classification EC* (pour *Enzyme Commission*). Bien que mise à jour lentement par rapport aux découvertes de nouvelles activités enzymatiques, la classification EC est largement utilisée pour décrire l'activité des enzymes et souvent, par extension, pour assigner une fonction enzymatique à un gène.

Nombre EC	Type d'enzyme	Type de réactions catalysées
1.-.-.-	Oxidoreductases	Réactions d'oxidoréduction
2.-.-.-	Transferases	Réactions de transfert de groupes fonctionnels
3.-.-.-	Hydrolases	Réactions d'hydrolyse d'un substrat en deux produits
4.-.-.-	Lyases	Réactions de coupure de liaisons covalentes par un procédé autre que l'oxydation ou l'hydrolyse
5.-.-.-	Isomerases	Réactions de réarrangement intramoléculaire, isomérisation
6.-.-.-	Ligases	Réactions de jonction covalente de deux molécules utilisant l'hydrolyse d'ATP

Tableau 1. Premier niveau de la classification EC. Un nombre EC se compose de quatre nombres représentant quatre niveaux de classification qui caractérisent de plus en plus finement la réaction catalysée. Le premier niveau présenté ici distingue six grandes classes de réactions. Le dernier niveau spécifie généralement les substrats précis de la réaction.

1.2.4 Cinétique des réactions métaboliques

Une bonne grandeur pour décrire le fonctionnement du métabolisme est la vitesse des réactions métaboliques, également appelée *flux*. En effet, la survie des cellules

[8] Voir http://www.chem.qmul.ac.uk/iubmb/

dépend fortement de leur capacité à transformer en permanence les métabolites pour produire l'énergie et construire ses constituants. Plus que les concentrations de tels ou tels métabolites, les flux des réactions renseignent directement sur les conversions métaboliques ayant lieu dans la cellule ; ils représentent en quelque sorte l'état fonctionnel du métabolisme. Nous reviendrons plus en profondeur sur la notion de flux et sa signification pour représenter l'état du métabolisme dans la section introduisant la modélisation.

La vitesse d'une réaction s'exprime généralement avec l'unité $mol.L^{-1}.s^{-1}$ qui décrit la quantité de substrats transformés par unité de volume de solution et par unité de temps[9]. Cette unité est bien adaptée à la description des flux lorsque les réactions se déroulent *in vitro,* mais l'est moins lorsqu'elle se déroulent dans des cellules ; on lui substitue alors l'unité $mmol.h^{-1}.(g\ DW)^{-1}$ où DW représente la masse sèche des cellules. Cette unité rapporte ainsi indirectement la quantité de substrat transformé par unité de temps à la quantité de cellule.

La vitesse d'une réaction enzymatique dépend de nombreux facteurs : concentration des substrats et produits, concentration de l'enzyme, efficacité catalytique de l'enzyme, température, pH, pression, entre autres facteurs… Sans vouloir exposer ici un état de l'art sur la cinétique enzymatique qui n'est pas le sujet de cette thèse, nous souhaitons rappeler au lecteur à titre illustratif un modèle simple de cinétique enzymatique qui permette d'appréhender l'influence de certains de ces facteurs. Michaelis et Menten déterminèrent, de manière d'abord empirique, une relation entre vitesse de réaction et concentration en substrat dépendant de deux paramètres liés à l'enzyme (Michaelis & Menten 1913; Cornish-Bowden 2004) :

$$v = v_{max} \frac{c_S}{K_m + c_S}$$

où v est la vitesse de réaction, c_S la concentration en substrat et v_{max} et K_m les deux paramètres en question. Le premier paramètre, v_{max}, représente la vitesse maximale que la réaction peut atteindre en présence d'une quantité fixe d'enzyme et pour un

[9] La vitesse de la réaction dépend de l'écriture de son équation bilan. La vitesse de production d'un produit (par la réaction) est en effet égale à la vitesse de la réaction multipliée par le coefficient stœchiométrique du produit dans l'équation bilan.

quantité saturante du substrat. Ce paramètre dépend linéairement de la quantité d'enzyme et traduit son efficacité à réaliser la transformation chimique. Définie en termes de nombre de molécules de substrat converties par une enzyme en une seconde, cette efficacité peut s'échelonner sur plusieurs ordres de grandeur, de 0.5 s^{-1} pour le lysozyme à 600 000 s^{-1} pour la carbonate déshydratase (Stephanopoulos et al. 1998; Barthelmes et al. 2007). Le second paramètre, K_m, également appelé constante de Michaelis, est égal à la concentration de substrat pour laquelle la vitesse de la réaction vaut ½ v_{max} (voir Figure 5). Ce paramètre est indépendant de la quantité d'enzyme et traduit l'affinité de l'enzyme au substrat (un K_m plus faible traduit une affinité plus élevée).

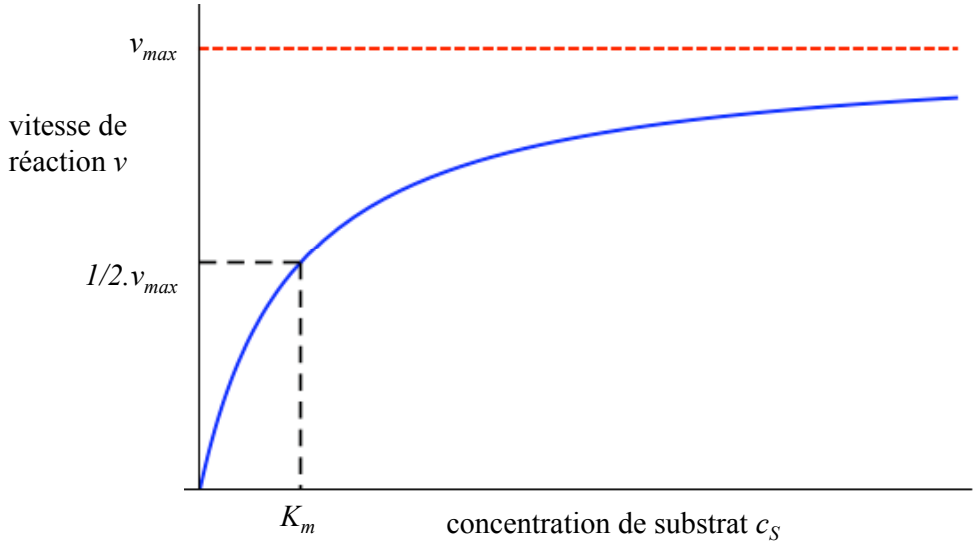


Figure 5. Relation entre vitesse de réaction et concentration de substrat pour une cinétique de type Michaelis-Menten.

Aux concentrations élevées de substrat ($c_S >> K_m$), la vitesse de la réaction tend vers v_{max}. L'enzyme est saturée et la vitesse de la réaction dépend linéairement de sa quantité. Aux concentrations faibles de substrat ($c_S << K_m$), la vitesse de la réaction tend vers $(v_{max}/K_m).c_S$ auquel cas elle dépend linéairement de la concentration en substrat et en enzyme. La concentration K_m délimite en quelque sorte les deux régimes de fonctionnement.

La cinétique de Michaelis-Menten s'interprète avec un modèle de transformation moléculaire simple où le substrat se lie réversiblement à l'enzyme avant d'être transformé irréversiblement en produit (Briggs & Haldane 1925) :

$$E + S \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} ES \overset{k_2}{\rightarrow} E + P$$

En supposant que le complexe enzyme-substrat *ES* est à l'état stationnaire, la relation de Michaelis-Menten est retrouvée avec :

$$v_{max} = k_2 c_{E,\text{tot}} \text{ et } K_m = \frac{k_{-1} + k_2}{k_1}$$

où $c_{E,\text{tot}}$ représente la quantité totale d'enzyme.

La cinétique de Michaelis-Menten traduit un mécanisme réactionnel relativement simple et en réalité beaucoup d'enzymes suivent des cinétiques bien plus complexes (Cornish-Bowden 2004). Elle est en revanche illustrative des influences respectives des enzymes et métabolites sur les flux de réaction et elle permet d'introduire les phénomènes de contrôle des réactions.

1.2.5 Contrôle des réactions métaboliques

Que la cinétique d'une réaction enzymatique soit Michaelienne ou non, l'enzyme en elle-même influence largement le flux de la réaction. Celui-ci dépend en effet à la fois de la quantité d'enzymes présentes et de leur efficacité à catalyser la réaction. Cette dépendance est exploitée de manière fondamentale par les organismes pour contrôler leur métabolisme, que ce soit simplement pour activer ou inactiver des réactions ou, de manière plus élaborée, pour ajuster finement la vitesse des réactions en fonction de leurs besoins. Les processus biologiques de contrôle sont généralement désignés sous le terme de *régulation métabolique*. On distingue typiquement deux grandes catégories de contrôles : (1) l'ajustement de la quantité d'enzymes et (2) la modulation directe de leur activité.

La régulation de la quantité d'enzymes s'opère communément en modulant les vitesses de production et de dégradation des enzymes (Stephanopoulos et al. 1998, pp.173-180). Des mécanismes complexes de régulation permettent en effet d'activer ou d'inhiber la transcription et la traduction de protéines en réponse à un signal particulier (par exemple la présence ou l'absence d'un métabolite particulier). Les microorganismes utilisent largement ce type de mécanisme, notamment pour adapter leur métabolisme aux environnements chimiques qu'ils rencontrent en ne produisant

que les enzymes appropriées à l'environnement. De même, en adressant les enzymes à des localisations spécifiques dans l'organisme, celui-ci peut contrôler à quel endroit les réactions se dérouleront, permettant par exemple d'éviter des interactions chimiques indésirables entre métabolites.

De nombreux mécanismes permettent également de réguler l'efficacité catalytique des enzymes. D'une part, les enzymes peuvent être totalement inactivées ou activées par des modifications covalentes irréversibles de leur structure ; ces modifications consistent fréquemment à les phosphoryler, ou à leur ajouter ou enlever divers groupes fonctionnels par l'intermédiaire de protéines ou de métabolites particuliers. D'autre part, et il s'agit de la classe de mécanisme la plus courante, des métabolites inhibiteurs ou activateurs peuvent interagir avec l'enzyme – souvent de manière réversible – pour modifier graduellement son activité. Divers mécanismes ont été identifiés, chacun conduisant à des comportements cinétiques souvent distinguables (Cornish-Bowden 2004). Ainsi, le métabolite régulateur peut tout aussi bien être un analogue du substrat et agir en tant que concurrent pour l'accès au site actif de l'enzyme, ou être différent et agir via un autre site sur la conformation de l'enzyme et altérer son efficacité catalytique ou son affinité au substrat (cas des enzymes allostériques).

Ces mécanismes de régulation agissent souvent de manière fine sur les flux des réactions en réponse à des signaux variés. Ceux-ci sont indispensables à l'organisme car ils lui permettent de réellement contrôler son « usine biochimique », pour notamment assurer la stabilité de sa composition chimique, économiser la production d'enzymes inutiles (en programmant par exemple leurs productions uniquement au moments opportuns (Zaslaver et al. 2004)) et répondre aux changements ou stimulus de leurs environnements (voire même les anticiper (Tagkopoulos et al. 2008; Mitchell et al. 2009) !).

1.2.6 Aspects thermodynamiques

Du point de vue thermodynamique, les organismes vivants sont des systèmes particuliers. Ils appartiennent à la classe des *systèmes dissipatifs* dont la caractéristique principale est de maintenir voire d'accroître leurs états d'ordre interne en prenant de l'énergie au milieu extérieur et lui retransmettant de l'entropie. Pour ce faire, ces systèmes doivent se maintenir en permanence hors de l'état d'équilibre

grâce à leurs échanges avec leur environnement ; ils sont fondamentalement ouverts[10] et tout arrêt de ces échanges conduit à leur disparition. Dans le cas des cellules vivantes, le maintien de cet état hors d'équilibre leur permet de croître et d'assurer la permanence de l'organisation de leur structure. Le métabolisme assure donc l'échange continuel de matière et d'énergie avec l'environnement : il exploite en général[11] l'énergie de métabolites d'énergie élevée et d'entropie faible importés de l'environnement en les transformant en métabolites d'énergie plus faible mais d'entropie plus élevée (von Stockar & Liu 1999; Stephanopoulos et al. 1998). De manière à assurer un flux de transformation permanent, qui est donc thermodynamiquement fondamental pour la vie de l'organisme, les réactions du métabolisme sont également elles-mêmes hors d'équilibre.

L'enthalpie libre de réaction, notée $\Delta_r G$, permet de décrire le sens d'évolution spontané des réactions. À température et pression constante, la réaction opère en effet dans le sens de diminution de l'enthalpie libre, tel que $\Delta_r G < 0$. Dès lors que $\Delta_r G$ atteint zéro, le flux net de la réaction devient nul. Le métabolisme doit ainsi s'assurer que les enthalpies libres des réactions sont bien négatives pour transformer les métabolites avec un flux net positif.

L'enthalpie libre de réaction dépend de l'enthalpie libre standard de réaction ($\Delta_r G°$) qui ne dépend que de la température et de la pression, et des concentrations de ses substrats et produits :

$$\Delta_r G = \Delta_r G° + RT.\ln(Q)$$

où R la constante des gaz parfaits et Q est le quotient de la réaction :

$$Q = \frac{p_1^{\gamma_{p1}}.p_2^{\gamma_{p2}}\ldots}{s_1^{\gamma_{s1}}.s_2^{\gamma_{s2}}\ldots}$$

[10] Un système ouvert peut échanger de l'énergie et de la matière avec le milieu extérieur, au contraire des systèmes isolés. Selon le second principe de la thermodynamique, un système isolé évolue toujours de manière à augmenter son entropie et tend invariablement à rejoindre son état d'équilibre.

[11] Dans le cas de la photosynthèse, l'énergie ne provient pas des métabolites mais de la lumière.

avec s_1, s_2 les activités[12] des substrats, p_1, p_2 celles des produits et les γ_i leurs coefficients stœchiométriques.

Deux « leviers » peuvent ainsi conduire à une enthalpie de réaction négative, le quotient de réaction et l'enthalpie libre standard de réaction. D'une part, le quotient de la réaction peut être diminué par un déséquilibre net de concentration dans lequel les substrats sont en excès par rapport aux produits. En consommant par exemple les produits au fur et à mesure de leur apparition, le métabolisme peut maintenir le déséquilibre de concentration et assurer la continuité de la réaction. Cependant, certaines conversions biochimiques possèdent des enthalpies libres de réaction trop élevées pour être favorisées uniquement par un déséquilibre de concentrations (en gardant des niveaux de concentrations « physiologiques »). Ceci est le cas par exemple de réactions de biosynthèse des constituants de la cellule, pour lesquelles les produits sont plus « énergétiques » que les substrats, conduisant à une enthalpie libre standard de réaction élevée. Ces réactions sont rendues réalisables en les couplant avec une réaction apportant de l'énergie, au premier rang desquelles figure l'hydrolyse de l'ATP. La réaction combinée, dont le couplage s'effectue d'ailleurs souvent au sein de la même enzyme (Stephanopoulos et al. 1998, pp.629-694), possède alors une enthalpie libre standard de réaction moins élevée la rendant thermodynamiquement réalisable aux concentrations physiologiques. Ce cas de figure illustre l'importance des cofacteurs énergétiques et des processus métaboliques associés à leur maintenance.

1.3 Structure et organisation du métabolisme

1.3.1 Le réseau métabolique

D'un point de vue plus global, le métabolisme d'un organisme se compose d'un nombre élevé de réactions (typiquement plus d'un millier de réactions distinctes pour une bactérie « de taille moyenne » comme *Escherichia coli* (Keseler et al. 2009)) qui

[12] Dans le cas des réactions en solution aqueuse, les activités s'identifient quasiment aux concentrations, moyennant quelques corrections liées notamment à la force ionique. On effectue également cette approximation dans le cas des réactions biochimiques intracellulaires, bien que le « solvant » constitué par le milieu cytoplasmique soit loin d'être aussi idéal qu'une solution aqueuse. Des corrections sont cependant souvent requises pour corriger les déviations trop importantes (Maskow & von Stockar 2005; Vojinović & von Stockar 2009).

convertissent un nombre tout aussi élevé de métabolites. Cependant, du fait que les métabolites sont partagés par les réactions (produits par certaines et consommés par d'autres), métabolites et réactions se structurent sous la forme d'un réseau, couramment appelé *réseau métabolique* (voir Figure 6). Au sein de ce réseau, on peut distinguer les enchaînements de réactions qui transforment étape par étape les métabolites, formant en quelques sortes des chemins de conversion dans le métabolisme. Comme mentionné ci-dessus à propos de la thermodynamique, l'enchaînement des réactions a d'ailleurs une réalité bien physique, du fait que pour maintenir les flux de conversion, les produits de chaque réaction doivent en permanence être réutilisés pour maintenir le déséquilibre thermodynamique. Cependant, une représentation complète du réseau métabolique telle que celle présentée sur la Figure 6 illustre uniquement de manière statique le métabolisme. Elle représente en effet l'ensemble des réactions chimiques pouvant avoir lieu, mais pas la réalité des conversions chimiques ayant lieu à un instant *t* dans la cellule. Toutes les conversions métaboliques possibles ne se réalisent pas toutes ensemble, mais plutôt en fonction des besoins de la cellule. Le contrôle des réactions métaboliques présenté ci-dessus joue à cet effet un rôle primordial pour orienter les conversions métaboliques selon certains chemins bien précis.

Comme illustré sur la Figure 6, certains métabolites sont connectés à un nombre de réactions nettement plus élevé que d'autres. Ceux-ci forment en quelque sorte des points d'embranchement[13] du réseau métabolique, à partir desquels commencent plusieurs branches métaboliques. En contraste, d'autres métabolites ne sont reliés simplement qu'à deux réactions, ne formant que des intermédiaires de voies de conversion. Nous verrons rapidement dans la partie suivante sur la modélisation que de nombreux travaux se sont attachés à étudier les propriétés topologiques des réseaux métaboliques.

[13] Le terme consacré en anglais, et parfois par abus en français, est « hub ».

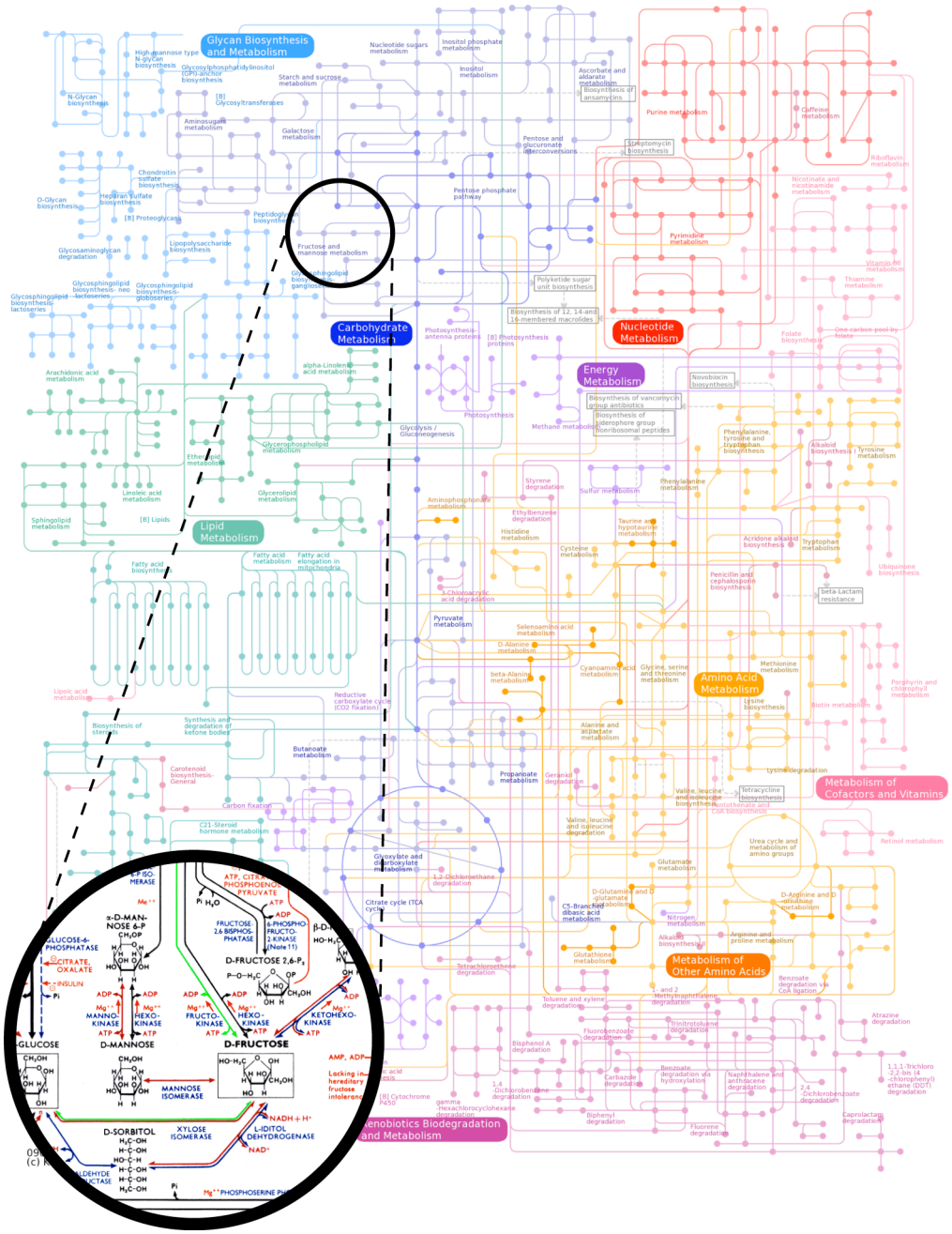


Figure 6. Illustration d'un réseau métabolique global. Les nœuds (points) correspondent à des métabolites et les liens (lignes) à des réactions (ou successions de réactions) convertissant les métabolites. Les grandes catégories fonctionnelles du métabolisme sont indiquées dans les encadrés. Dans le cercle : aperçu détaillé d'une partie du réseau métabolique. Source des cartes : KEGG (http://www.genome.jp/kegg/atlas/) et Roche Applied Science (http://www.expasy.ch/tools/pathways/).

De manière à obtenir une description fonctionnelle claire du métabolisme, les biochimistes ont traditionnellement regroupé les réactions en *voies métaboliques*, qui peuvent être vues comme des parties du réseau métabolique – souvent des enchaînements linéaires de réactions. La grande majorité des voies métaboliques sont définies pour correspondre à des fonctions métaboliques bien précises, mais cette classification est arbitraire et varie souvent d'une personne à l'autre, reposant parfois

sur des critères historiques relatifs au contexte de leur découverte, organisationnels ou purement subjectifs. Le découpage du réseau métabolique en voies fournit cependant une vision modulaire du métabolisme qui a permis d'en appréhender le fonctionnement global d'une manière simple.

1.3.2 Organisation globale du métabolisme

Sans chercher à rentrer dans le détail des voies composant le métabolisme, celui-ci peut être décrit de manière simple en distinguant une partie *catabolique* et une partie *anabolique*.

Les réactions du métabolisme catabolique ont pour fonction de dégrader (cataboliser) des composés chimiques provenant principalement de l'environnement de l'organisme[14] pour fournir de l'énergie, des cofacteurs réducteurs et des *précurseurs* nécessaires à la synthèse des constituants de la cellule. On peut y distinguer également deux types de voies en fonction de « l'originalité » des métabolites à cataboliser. D'une part des voies relativement ubiquitaires entre les organismes sont en charge de produire massivement l'énergie de la cellule ainsi que les précurseurs et cofacteurs à partir de métabolites communs ; elles sont parfois désignées par le terme *métabolisme central*. Parmi les voies métaboliques appartenant à cette première catégorie figurent notamment la glycolyse (produisant de l'énergie, des cofacteurs réduits et des précurseurs par oxydation d'hexoses, généralement glucose et fructose), le cycle de Krebs (ou cycle de l'acide citrique, voie métabolique centrale produisant de l'énergie, des cofacteurs réduits et des précurseurs par oxydation de l'acide citrique (voir Figure 7)), la phosphorylation oxydative (ou respiration, produisant de l'énergie par oxydation des cofacteurs réduit généralement grâce à l'oxygène du milieu) et des voies de fermentation (permettant de générer de l'énergie et de réoxyder les cofacteurs réduits en milieu anaérobie).

[14] le catabolisme peut également recycler des métabolites internes à la cellule.

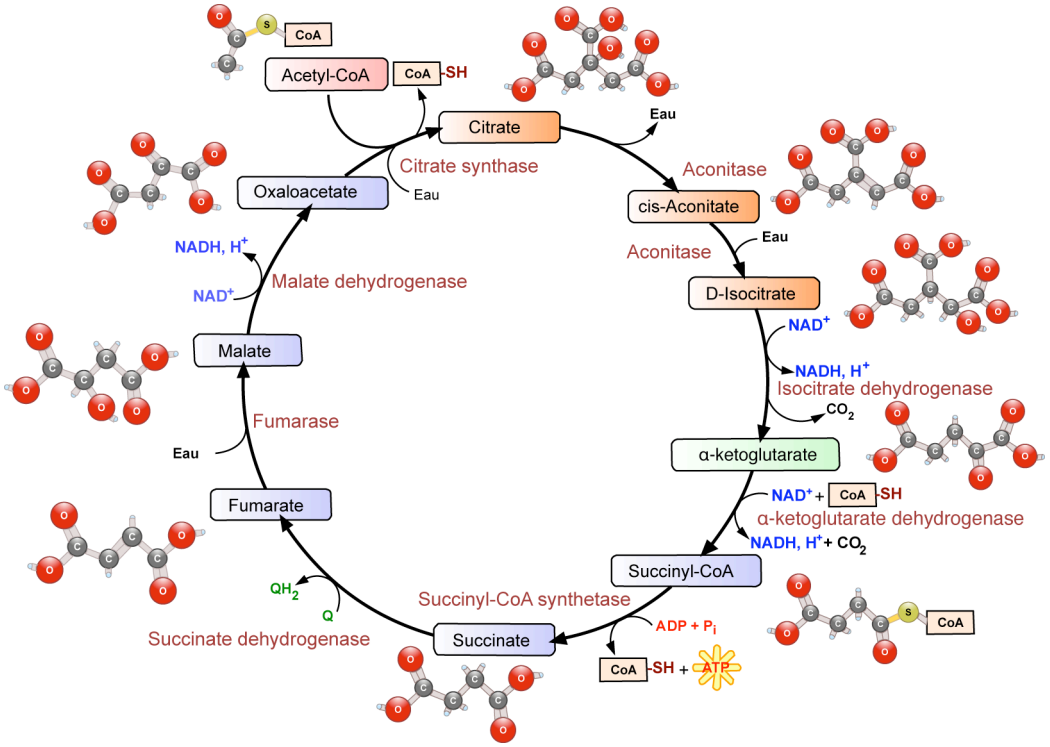


Figure 7. Schéma du cycle de l'acide citrique (citrate), pour *E. coli*. L'acetyl-coA est condensé avec l'oxaloacetate pour former le citrate. Celui-ci est progressivement oxydé et perd deux carbones sous forme de CO_2. Au cours d'un cycle, 1 ATP est produit, ainsi que 3 NADH et 1 QH_2 (ubiquinol), potentielles sources d'ATP si oxydées par l'oxygène dans la chaîne respiratoire. Adapté de Wikipedia (http://en.wikipedia.org/wiki/Citric_acid_cycle)

Lorsque les métabolites du métabolisme central ne sont pas directement disponibles dans l'environnement, ce qui est en réalité majoritairement le cas, des voies plus spécialisées se chargent de les créer en dégradant les composés qui s'y trouvent, tout en collectant l'énergie issue de cette dégradation. La répartition de ces voies dans les organismes est beaucoup plus disparate car elle dépend fortement de leurs milieux habituels de vie. Ainsi, les entérobactéries possèdent des voies de dégradation spécialisées dans l'utilisation de nombreux sucres tandis que les bactéries du sol disposent plutôt de voies de dégradation de composés issus des plantes, comportant par exemple des cycles aromatiques. Alors que le panel de composés « catabolisables » est extrêmement vaste, les voies de dégradation correspondantes ont cependant en commun de produire *in fine* des métabolites communs (centraux) pouvant être réutilisés ensuite par le reste du réseau métabolique.

Le métabolisme anabolique synthétise quant à lui les constituants de la cellule à partir des précurseurs et cofacteurs créés par le catabolisme ou directement importés de l'extérieur. Les cellules sont en effet constituées d'un assemblage hétérogène de macromolécules aux fonctions nombreuses (voir Tableau 2) , notamment le maintien

de la structure de la cellule (lipides, peptidoglycane), la conservation et la transmission d'information (ARN, ADN, protéines), la catalyse des réactions biochimiques (protéines, ARN). Ces molécules sont nommées macromolécules car celles-ci sont des molécules de grande taille formés en général par polymérisation de métabolites élémentaires, par exemple les acides aminés pour les protéines, les acides nucléiques pour l'ARN et l'ARN, et l'acetyl-coA pour les lipides. Les molécules synthétisées par l'organisme ne servent pas toutes directement au fonctionnement de la cellule, mais peuvent être utilisées également par l'organisme pour interagir avec son environnement ou d'autres organismes. Certaines peuvent par exemple être excrétées pour solubiliser l'environnement extérieur, aider à se sédentariser sur un support (création de biofilms), ou éliminer des bactéries concurrentes (synthèse d'antibiotiques). En résumé, le métabolisme anabolique comprend l'ensemble des réactions permettant à l'organisme de créer l'arsenal de composés chimiques qui lui seront utiles.

Macromolécule	**Fraction de la masse sèche totale**
Protéine	55,0%
ARN	20,5%
ARNr	16,7%
ARNt	3,0%
ARNm	0,8%
ADN	3,1%
Lipide	9,1%
Lipopolysaccharide	3,4%
Peptidoglycane	2,5%
Glycogène	2,5%
Fraction soluble	3,9%

Tableau 2. Composition moyenne en macromolécules de cellules d'*Escherichia coli*. Données issues de Neidhardt et Umbarger (1996).

Les précurseurs utilisés par le métabolisme anabolique sont quasiment universels et appartiennent au métabolisme central. Cependant, comme mentionné plus haut, certains organismes ne peuvent pas synthétiser par eux-mêmes tous les métabolites requis à leur survie et doivent importer ceux-ci de leur environnement. D'autres organismes sont au contraire extrêmement efficaces pour les synthétiser à partir de molécules très basiques comme des composés à un seul carbone (CO_2 ou CH_4). De même que pour les voies cataboliques, une partie des voies anaboliques est relativement universelle entre les organismes – ce sont celles aboutissant aux

constituants ubiquitaires et vitaux des cellules. L'autre partie des voies anaboliques est, au contraire, répartie très inégalement entre les organismes et brosse un éventail beaucoup plus grand de molécules biologiques. En particulier, on trouve une très grande variété de molécules issues du métabolisme secondaire chez les organismes supérieurs et notamment celui des plantes qui ont développé un vaste arsenal de défense « moléculaire ». Cependant, cette thèse étant focalisée sur le métabolisme des microorganismes, nous ne ferons que l'évoquer occasionnellement.

1.4 Méthodes d'exploration du métabolisme

1.4.1 Élucidation expérimentale des voies métaboliques

Historiquement, l'identification des voies métaboliques débuta peu après la découverte des premières enzymes à la toute fin du 19^{e} siècle. Le développement de méthodes de purification et de caractérisation des enzymes associées aux techniques d'identification chimique et de marquage radioisotopique des métabolites ainsi qu'à l'étude de la physiologie des microorganismes permit alors rapidement à un grand nombre de biochimistes de découvrir les principales voies métaboliques de divers organismes modèles. Ce travail, qui fut notamment très prononcé au milieu du 20^{e} siècle, généra une connaissance considérable sur l'enchaînement des réactions dans les voies métaboliques mais aussi sur les caractéristiques catalytiques, cinétiques et régulatoires des enzymes. La classification EC de l'IUBMB (voir 1.2.3) fut d'ailleurs créée à la fin des années 50 pour fournir une classification et une nomenclature uniformisée des enzymes, permettant d'organiser les nombreuses activités enzymatiques déjà identifiées. Dans un deuxième temps (deuxième moitié du 20^{e} siècle), le développement des techniques de biologie moléculaire permit d'associer des gènes aux enzymes identifiées et apportèrent un angle de vision complémentaire sur le fonctionnement du métabolisme.

La somme des connaissances accumulées sur le métabolisme à la fin du 20^{e} siècle fut considérable et relativement complète pour quelques organismes modèles, au premier rang desquels *Escherichia coli* pour les procaryotes (Neidhardt 1996) et *Saccharomyces cerevisiae* pour les eucaryotes (Strathern et al. 1982). Pour ces organismes modèles, la majeure partie des voies métaboliques fut décrite en détail, permettant aujourd'hui d'appréhender la globalité de leur métabolisme. Cependant, cette connaissance fut accumulée sous la forme d'un nombre tout aussi considérable

de publications. Pour organiser ces informations disparates, des initiatives de regroupement de l'information au sein de bases de données ont vu le jour depuis une vingtaine d'années. Deux grands types de bases de données liées au métabolisme ont été développés. D'une part, des répertoires de données biochimiques sur les enzymes ; les ressources les plus significatives sont BRENDA (Barthelmes et al. 2007) et ENZYME (Bairoch 2000). D'autre part, des bases de données répertoriant selon diverses organisations la biochimie des voies métaboliques. Parmi ces dernières, KEGG (Kanehisa et al. 2007) et MetaCyc (Caspi et al. 2008) regroupent les voies indépendamment des organismes, EcoCyc est spécifique au métabolisme d'*E. coli* (Keseler et al. 2009), et UM-BBD se concentre sur les voies de dégradation microbiennes (Ellis et al. 2006).

Malgré le développement de techniques d'investigation en biochimie de plus en plus performantes – incluant notamment la chromatographie, la RMN et la spectrométrie de masse – la découverte de nouvelles activités enzymatiques marqua le pas à la fin du 20^{e} siècle. D'une part, la majeure partie des voies métaboliques principales des organismes modèles cultivables fut déjà élucidée, et d'autre part, les centres d'intérêts majoritaires en biologie se sont déplacés plutôt sur des sujets de biologie moléculaire.

Aujourd'hui, la recherche de nouvelles activités enzymatiques bénéficie cependant d'un regain d'intérêt notable. Tout d'abord, l'augmentation rapide du nombre de génomes et de métagénomes[15] séquencés identifie une quantité toujours plus importante de gènes de fonctions inconnues, dont une fraction significative code vraisemblablement pour des enzymes métaboliques. Inversement, l'étude des phénotypes de croissance (voir ci-dessous, chapitre 2), de la physiologie des microorganismes et du contenu métabolomique[16] des cellules (Dunn et al. 2005; Breitling et al. 2008; Steuer 2006) permet d'identifier, parfois de manière ciblée, des

[15] On désigne par métagénome tout matériel génétique d'une communauté microbienne extrait directement d'un échantillon. Il contient ainsi indistinctement les matériels génétiques des divers organismes présents dans cet environnement, permettant leur étude en s'affranchissant de l'isolement des organismes et de leur mise en culture, souvent difficiles voire impossibles.

[16] Le métabolome, dans la lignée des termes en -ome (p.ex. génome, transcriptome, protéome), désigne l'ensemble des métabolites présents dans une cellule.

activités métaboliques dont les gènes et enzymes sont inconnus[17] (Lespinet & Labedan 2006b; Lespinet & Labedan 2006a; Pouliot & Karp 2007). De nombreux projets se sont développés récemment pour tenter de relier gènes et activités enzymatiques sur ces deux bases. Les résultats attendus sont prometteurs, dans la mesure où la disponibilité du matériel génétique pour de nombreux organismes, qu'ils soient cultivables ou non, voire non identifiés à l'instar des métagénomes, ouvre la voie à l'étude de l'activité d'enzymes auparavant difficilement accessibles. Les méthodes expérimentales mises en œuvre incluent notamment le criblage d'activité de banques d'enzymes (Kitagawa et al. 2005) sur des ensembles de substrats (Saghatelian et al. 2004; Saito et al. 2006) et la recherche d'associations entre gènes et phénotypes métaboliques (Aghaie et al. 2008).

1.4.2 Méthodes bioinformatiques de reconstruction des réseaux métaboliques

La mise en évidence expérimentale des activités métaboliques opérant dans un organisme constitue la preuve la plus directe de leur existence. Cependant, quand bien même le débit des techniques expérimentales correspondantes a fortement augmenté ces dernières années, celles-ci restent encore beaucoup trop lourdes à réaliser pour élucider globalement le métabolisme de tout nouvel organisme.

La possibilité de séquencer des génomes complets à moindre coût offre aujourd'hui une solution alternative efficace (Feist et al. 2009). En effet, le développement du séquençage s'est accompagné de la mise en place de méthodes bioinformatiques permettant d'une part de détecter les gènes sur la séquence du génome et d'autre part d'inférer leurs fonctions, processus appelé *annotation* du génome (Médigue & Moszer 2007). L'inférence de la fonction des gènes se base essentiellement sur la recherche d'homologies avec les gènes de fonctions déjà connues : deux gènes codant pour des séquences protéiques très proches ont de fortes chances de coder pour des protéines de fonctions similaires. De cette manière, les activités enzymatiques associées à certains gènes peuvent être propagées par homologie aux gènes nouvellement séquencés, bien que la transitivité de ce processus puisse induire des erreurs d'annotation. Afin d'augmenter la fiabilité des annotations

[17] Activités orphelines

prédites, les processus actuels d'annotation automatique combinent les sources d'informations (Médigue & Moszer 2007). En particulier, ceux-ci étudient par exemple le contexte génomique[18] des gènes pour confirmer et préciser les annotations prédites. En moyenne, ces méthodes permettent de prédire une fonction pour 50 à 80 pour cent des gènes d'un organisme bactérien nouvellement séquencé (Serres et al. 2004).

Les fonctions enzymatiques prédites par les méthodes d'annotation constituent une source primordiale de données permettant de reconstruire le réseau métabolique de l'organisme étudié. Les méthodes classiquement utilisées pour annoter les génomes sont cependant généralistes et ne précisent pas nécessairement le détail des conversions métaboliques associées à une fonction enzymatique. De plus, la spécificité des conversions catalysées par une enzyme prédite peut se révéler difficile à établir sur la seule base d'homologie de séquences. Des méthodes dédiées à la reconstruction du métabolisme à partir de génomes annotés ont été développées pour répondre à ces faiblesses. Elles reposent sur des bases de données de réactions métaboliques qui leur permettent d'énumérer les réactions potentiellement catalysées par les fonctions enzymatiques annotées et d'en détailler la biochimie. Afin de sélectionner les réactions les plus probables et de préciser leur spécificité, ces méthodes examinent également leur contexte métabolique : l'existence d'une réaction prédite peut en effet être confortée par la présence d'autres réactions impliquant ses substrats et ses produits.

D'autres méthodes bioinformatiques ont été élaborées pour détecter les activités métaboliques manquant dans le réseau métabolique reconstruit. Une partie d'entre-elles se base sur la connaissance des voies métaboliques complètes pour détecter les « trous » dans les voies constitués par les réactions manquantes. De nombreuses méthodes ont également été développées pour combler ces trous et rechercher des

[18] On appelle *contexte génomique* d'un gène toute information apportée par son voisinage chromosomique. Il peut s'agir par exemple d'un type de fonction biologique lorsque plusieurs gènes voisins possèdent des rôles contribuant à une fonction biologique particulière. L'utilisation du contexte génomique peut être renforcée par la recherche de *synténies*, c.-à-d. de groupes de gènes voisins co-conservés chez différents organismes. La conservation groupée des gènes peut être un indice du fait que les gènes contribuent ensemble à une fonction biologique.

gènes candidats, sur la base de leur contexte génomique, de leur occurrence phylogénétique ou de leur expression.

Nous n'avons énuméré ici que les principales idées des méthodes bioinformatiques de reconstruction des réseaux métaboliques. Nous en effectuerons un état de l'art beaucoup plus détaillé plus loin dans le manuscrit, dans la revue consacrée aux modèles globaux du métabolisme (voir section 3.2.1).

1.4.3 Vers une étude globale du métabolisme

L'efficacité des méthodes bioinformatiques de reconstruction du métabolisme dépend directement de la variété de voies métaboliques et d'activités biochimiques préalablement élucidées et de leur « proximité » avec l'organisme étudié. L'accumulation considérable de connaissances sur les voies métaboliques fait qu'aujourd'hui une part significative du métabolisme d'un organisme nouvellement séquencé peut être reconstruite à l'aide de ces méthodes. À l'instar des organismes modèles pour lesquels la majeure partie du métabolisme fut découverte par expérimentation, la reconstruction *in silico* du métabolisme des organismes dont on dispose de la séquence ouvre la voie à l'exploration globale de leurs voies métaboliques et de leurs capacités de conversions. Il est évident que ces méthodes ne peuvent détecter *de novo* des fonctions métaboliques originales, ces dernières n'ayant jamais été identifiées auparavant et encore moins associées à un gène. Cependant, en reconstituant rapidement la part déjà connue du métabolisme, ces méthodes contribuent justement à en cerner la partie encore inconnue qui constitue souvent le cœur d'intérêt de l'étude.

Le choix d'étudier le métabolisme d'un organisme en particulier est, dans de nombreux cas, guidé par une caractéristique de sa physiologie : par exemple sa capacité à exploiter un nutriment particulier, son efficacité accrue à survivre dans un environnement donné ou son aptitude à produire un métabolite. Ces observations traduisent généralement à l'échelle cellulaire des caractéristiques particulières de leur métabolisme, caractéristiques dont l'élucidation est le but de ces études. De manière plus générale, effectuer le lien entre des observations macroscopiques du métabolisme d'une cellule – efficacité de conversion, vitesses de croissance ou de consommation/production de métabolites, capacité de survie dans des environnements chimiques donnés – et le détail des activités enzymatiques identifiées présente de

nombreux intérêts. D'une part, ces observations macroscopiques apportent des informations supplémentaires sur le métabolisme, permettant d'évaluer la pertinence des voies métaboliques reconstruites et de guider leur investigation. D'autre part, ces observations sont la trace du fonctionnement *in vivo* du métabolisme et des flux de conversions ayant réellement lieu dans la cellule. Elles complètent avantageusement la vision statique des voies métaboliques en donnant des indications sur les conversions réellement à l'œuvre.

Les techniques expérimentales d'investigation des « états cellulaires » ont en outre considérablement progressées dans la dernière décennie, à la fois dans leurs précisions et leurs débits (Joyce & Palsson 2006). Elles permettent ainsi d'étudier à grande échelle le niveau de transcription des gènes (*transcriptomique*), la concentration intracellulaire des protéines (*protéomique*), des métabolites (*métabolomique*) et, pour l'instant dans une moindre mesure, le niveau des flux des réactions métaboliques (*fluxomique*). Les données générées fournissent des indications directes sur l'état des acteurs du métabolisme, offrant la capacité sans précédent d'accéder aux états physiologiques internes de la cellule. Néanmoins, elles nécessitent d'être intégrées, interprétées et combinées dans le contexte global du métabolisme pour en tirer des conclusions pertinentes sur le fonctionnement biochimique réel de la cellule.

La connaissance à grande échelle du métabolisme offre justement la possibilité d'explorer globalement le fonctionnement du métabolisme et de le relier aux comportements macroscopiques observés.

Afin de réduire la complexité des réseaux métaboliques, les biochimistes et les microbiologistes ont traditionnellement utilisé le regroupement des réactions en voies métaboliques pour raisonner globalement sur les conversions métaboliques (voir 1.3.1). Chaque voie métabolique y est vue indépendamment l'une de l'autre et est caractérisée par son bilan, à savoir la transformation des métabolites d'entrées en métabolites produits. En raisonnant sur les conversions des quelques métabolites clés par ces voies métaboliques, un aperçu global de la physiologie de la cellule pouvait alors être prédit et corroboré avec les observations réelles. De même, ce découpage du métabolisme est régulièrement utilisé pour visualiser les données expérimentales dans

le contexte du métabolisme global (Kanehisa et al. 2006; Paley & Karp 2006; Shannon et al. 2003).

Cette méthode de raisonnement trouve toutefois rapidement ses limites, pour au moins deux raisons. Premièrement, la juxtaposition de voies métaboliques simplifie souvent de manière exagérée les interconnexions existant entre processus métaboliques. En effet, le bon fonctionnement d'une voie métabolique requiert fréquemment l'exécution de conversions métaboliques « annexes », par exemple la synthèse de précurseurs particuliers ou la régénération de métabolites cofacteurs. Un raisonnement basé uniquement sur l'étude de l'enchaînement des voies métaboliques risque ainsi de laisser de côté certaines interdépendances métaboliques jouant un rôle significatif dans le comportement global. Ensuite, l'étude du fonctionnement réel du métabolisme nécessite dans un grand nombre de cas de tenir compte de ses aspects quantitatifs : comment se répartissent les flux de matière dans les voies métaboliques, quelle quantité d'énergie est consommée par le fonctionnement de ces voies ? Quand bien même il est possible de prendre manuellement en considération ces aspects quantitatifs pour un nombre limité de voies métaboliques, étendre leur usage à l'échelle du métabolisme entier nécessite l'emploi de méthodes plus systématiques.

Les modèles mathématiques du métabolisme répondent justement à ces deux types de difficultés (voir section 3). Ils combinent généralement une description plus ou moins détaillée des activités métaboliques présentes dans la cellule avec la capacité de raisonner sur leurs états fonctionnels (concentrations de métabolites et d'enzymes, flux de réactions) en appliquant les principes physiques déterminants. Ils ont ainsi été particulièrement utilisés pour étudier la dynamique précise de voies métaboliques, intégrer des données métaboliques expérimentales de diverses origines et prédire des comportements métaboliques macroscopiques. Nous effectuerons une revue plus complète des types de modélisation métabolique existant dans la section 3.

Notre thèse s'inscrit directement dans ce schéma. Son objectif est de développer des outils et méthodes permettant au mieux d'interpréter un certain type d'observations métaboliques macroscopiques – les phénotypes de croissance (voir section 2) – à la lumière du réseau métabolique, en utilisant pour cela la modélisation mathématique.

2 Phénotypes de croissance et essentialité de gènes

2.1 Phénotypes de croissance

On appelle *phénotype* toute caractéristique observable d'un organisme. Un *phénotype de croissance* désigne ainsi toute caractéristique propre à la croissance des microorganismes. Par exemple : dans quels environnements sont-ils capables de se développer, à quelle vitesse ; dans quelles proportions les nutriments sont-ils consommés, quels sont les composés produits. Alors que les approches d'exploration du métabolisme présentées ci-dessus sont particulièrement adaptées pour décrire le détail des conversions chimiques à l'œuvre, l'étude des phénotypes de croissance fournit des informations d'échelle plus large mais néanmoins complémentaires sur le fonctionnement du métabolisme.

Les expériences de cultures suivies de microorganismes permettent de mesurer à intervalles de temps réguliers la composition chimique de l'environnement des organismes, ainsi que la quantité de biomasse créée (voir Figure 8). À l'aide de ces mesures, les échanges métaboliques entre les organismes et leur environnement (consommation de substrats, excrétion de produits) peuvent être déterminés quantitativement et reliés à leur vitesse de croissance. Ces observations de la physiologie des organismes fournissent des informations importantes sur le fonctionnement *in vivo* du métabolisme, quand bien même elles sont d'échelle macroscopique. Par exemple, les suivis de la consommation d'oxygène et de substrat carboné ainsi que de la production de dioxyde de carbone sont traditionnellement utilisés pour évaluer le rendement de production énergétique des microorganismes (Neijssel et al. 1996). De même, lorsque l'ensemble des échanges suivis est suffisamment exhaustif, un bilan « d'utilisation du carbone » par le métabolisme peut être effectué, permettant de déterminer quel usage est fait des nutriments carbonés par les organismes. La répartition du carbone entre les molécules de dioxyde de carbone, produits de fermentation et biomasse fournit des indications quant au régime métabolique en cours dans les microorganismes.

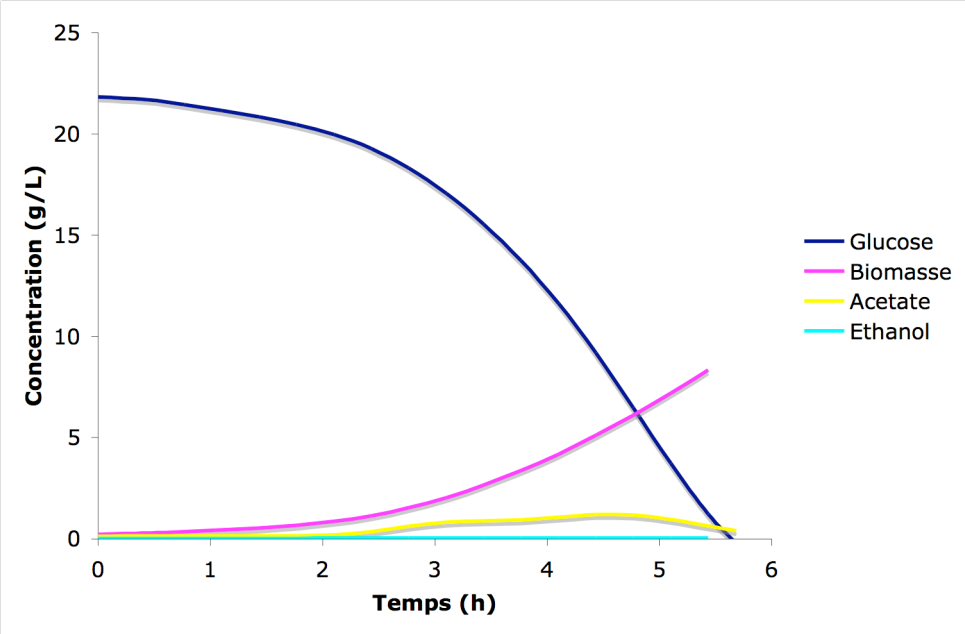


Figure 8. Expérience de croissance suivie pour une population d'*Escherichia coli* cultivée dans un fermenteur en condition aérobie. Tracé des concentrations de glucose, biomasse, acétate et éthanol en fonction du temps.

Une autre classe d'expériences consiste à cribler à grande échelle la croissance des microorganismes sur des milieux distincts. En utilisant des milieux minimaux[19] dans lesquels les métabolites sources de carbone, azote, phosphore et soufre sont testés de manière systématique, ces expériences permettent de déterminer rapidement quels métabolites sont exploités par l'organisme. Ces résultats sont utilisés de manière courante par les microbiologistes pour classer les microorganismes et définir les espèces[20] ; ils fournissent également des indications quant à leurs environnements naturels. De plus, le fait d'associer les métabolites aux types de contributions métaboliques (apport en carbone, azote, phosphore ou soufre ; accepteur d'électron) guide la recherche des voies métaboliques sous-jacentes.

Habituellement, ces expériences sont réalisées « manuellement » en testant la croissance sur un ensemble de milieux minimaux préparés séparément. Cependant, la société Biolog a récemment développé et commercialisé un procédé permettant d'augmenter le débit de ces tests en utilisant des plaques à 96 puits contenant des

[19] Un milieu minimal est un milieu de culture de composition contrôlée, couvrant de manière minimale les besoins en nutriments de la cellule. Généralement, un milieu minimal possède un seul type de métabolite contribuant à chaque apport de carbone, azote, phosphore et soufre.
[20] Voir http://www.bacterio.cict.fr/

milieux minimaux distincts. Ces milieux sont tous des variations autour d'une même base, permettant de cribler de manière systématique les sources de carbone, azote, soufre ou phosphore (Bochner 2009). Après inoculation, la croissance et l'activité métabolique[21] sont automatiquement suivies au cours du temps dans chacun des puits (voir Figure 9). À l'heure actuelle, Biolog propose 20 plaques de phénotypage différentes, représentant un ensemble de 1920 milieux. Parmi eux, 190 testent des sources de carbone, 380 des sources d'azote et 95 des sources de soufre et de phosphore. Les milieux restants évaluent la sensibilité des cellules à diverses molécules chimiques, dont une majorité d'antibiotiques, ainsi qu'à des changements de pH et de force ionique.

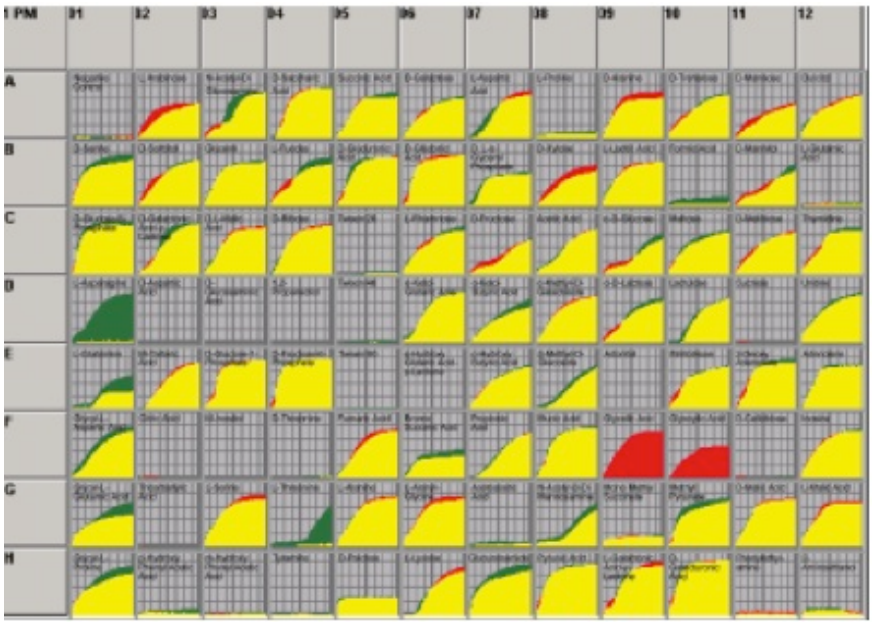


Figure 9. Activités métaboliques comparées de deux souches bactériennes sur 96 sources de carbone distinctes. En rouge et vert, cinétiques de croissance spécifiques à chacune des souches. En jaune, partie commune de leurs cinétiques. Extrait de Bochner (2009)

2.2 Exploration génétique des phénotypes de croissance

Les phénotypes représentent des caractéristiques de l'organisme étudié, qui sont souvent aisément observables. Le développement de la biologie moléculaire, et notamment du génie génétique, a fourni des outils permettant d'investiguer ces phénotypes et de rechercher des associations fonctionnelles entre eux et les gènes. Une grande classe d'expériences développées dans cet esprit consiste à rechercher les

[21] Dans la méthodologie Biolog, un rapporteur coloré de la respiration cellulaire (le tetrazolium) est incorporé aux puits. Son niveau est suivi en parallèle à la quantité de cellule obtenue par mesure de la densité optique (Bochner 2009).

pertes de phénotypes provoquées par la suppression d'un gène. Pour ce faire, elles comparent les phénotypes de l'organisme sauvage avec ceux de l'organisme dont le gène ciblé a été inactivé ou éliminé par génie génétique, organisme *mutant*. Ces expériences permettent ainsi de mettre expérimentalement en évidence des liens de causalité entre la présence d'un gène et l'occurrence d'un phénotype particulier.

S'agissant des phénotypes de croissance, l'altération recherchée la plus significative est la létalité, c'est-à-dire la perte de la capacité à croître consécutivement à l'inactivation ou l'élimination d'un gène. Ces gènes sont généralement désignés comme *essentiels*[22]. De plus, la létalité peut n'être observée que pour une partie des environnements testés, on parle dans ce cas de gènes *conditionnellement essentiels.*

Dans cette section, nous donnerons un bref aperçu des principales techniques expérimentales permettant de tester à grande échelle les effets de l'inactivation de gènes, puis nous présenterons les principales applications de ces méthodes, en particulier pour l'exploration du métabolisme.

2.2.1 Techniques expérimentales

Deux aspects de ces techniques expérimentales peuvent être distingués (bien qu'ils ne soient pas complètement indépendants) : d'une part la méthode utilisée pour inactiver ou éliminer les gènes et d'autre part la stratégie employée pour détecter à grande échelle les gènes essentiels.

Inactivation des gènes

Trois catégories de techniques expérimentales permettent d'inactiver les gènes : (1) l'insertion d'un transposon dans le gène, (2) la délétion totale du gène par recombinaison et (3) l'inactivation des transcrits par ARN interférents.

Le mécanisme de transposition[23] offre un outil d'inactivation génique puissant et relativement simple à utiliser (Hayes 2003). Le principe de l'inactivation consiste à

[22] Par opposition aux gènes *non-essentiels*. La terminologie principalement utilisée en anglais est « *essential genes* » et « *dispensable genes* ».
[23] Les *transposons* sont des séquences ADN pouvant se déplacer, et a fortiori, s'intégrer de manière autonome dans le génome. Le mécanisme de *transposition*

insérer le transposon au sein de la séquence du gène ciblé de manière à empêcher la transcription de sa séquence complète (voir Figure 10). Les sites d'insertion des transposons étant difficilement contrôlables et modifiables en fonction des gènes ciblés, les techniques d'inactivation génique emploient des stratégies d'insertion aléatoire des transposons dans le génome. D'un point de vue pratique, les transposons utilisés sont donc choisis pour pouvoir s'insérer de la manière la moins biaisée possible à n'importe quel endroit du génome. Différentes stratégies expérimentales ont été développées pour favoriser la transposition ; d'une part des stratégies *in vivo* utilisant des plasmides ou des phages introduisant les séquences ADN des transposons dans les cellules, et d'autre part des stratégies *in vitro*, réalisant tout ou partie de la transposition hors de la cellule avant intégration dans le génome (Reznikoff & Winterberg 2008). Les techniques d'inactivation par transposition présentent l'avantage de pouvoir « inactiver » très facilement de nombreux sites dans le génome, de manière non ciblée. Combinées avec des méthodes efficaces de sélection des mutants (voir ci-dessous), ces techniques permettent de révéler rapidement des altérations chromosomiques délétères. L'interprétation de « l'altération chromosomique » provoquée par l'insertion d'un transposon n'est cependant pas forcément évidente. D'une part, l'inactivation du gène par insertion n'est en effet pas forcément réalisée et, d'autre part, l'insertion peut provoquer des effets polaires perturbant la transcription de gènes éloignés du site d'insertion, mais présents dans le même opéron. D'autre part, le biais d'insertion des transposons, même faible, rend inaccessibles à l'étude certaines régions du génome et perturbe l'analyse statistique des études par insertion aléatoire (Hayes 2003).

---

repose sur l'utilisation d'une enzyme, la *transposase*, capable d'exciser puis d'intégrer le transposon dans la séquence ADN.

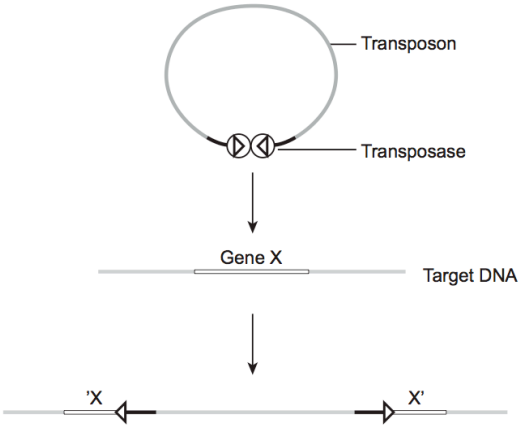


Figure 10. Illustration du mode d'action d'un transposon. Le transposon, préalablement excisé de son vecteur et attaché à la transposase, se lie au gène cible. La transposase catalyse alors l'intégration du transposon dans la séquence de l'ADN ciblé, altérant la structure du gène X. Illustration issue de Reznikoff & Winterberg (2008).

Afin de maîtriser au mieux l'inactivation des gènes, l'excision précise et totale du gène de la séquence génomique est parfois privilégiée, même si le coût humain et matériel est nettement supérieur à celui des méthodes de transposition. Les techniques utilisées à cet effet s'appuient généralement sur les processus de recombinaison homologue permettant de remplacer la région génomique ciblée par une séquence de substitution introduite dans la cellule, portant un marqueur de sélection destiné à identifier les mutants corrects (de Berardinis et al. 2008; Baba et al. 2006; Giaever et al. 2002). La séquence de substitution peut ensuite être éliminée pour réduire les effets polaires et pouvoir répéter le processus de délétion à un autre endroit du génome. Nous détaillerons plus loin dans ce manuscrit un exemple de protocole de délétion utilisé pour la bactérie *Acinetobacter baylyi* (section 4.3). Bien que très précise, chaque délétion doit cependant être réalisée individuellement, rendant le processus laborieux pour la délétion systématique de nombreux gènes (du fait du nombre important de constructions génétiques spécifiques à réaliser).

Enfin, un dernier grand type de technique utilisé consiste à inactiver, non pas le gène directement, mais ses transcrits à l'aide d'ARN interférents. Largement utilisée dans le cas des cellules eucaryotes et notamment d'organismes supérieurs (Dykxhoorn et al. 2003), cette technique est employée également avec succès chez les bactéries (Engdahl et al. 1997; Ji et al. 2001; Forsyth et al. 2002). Elle permet d'inactiver spécifiquement l'action du gène ciblé avec des risques moindres d'interactions avec celles d'autres gènes. L'inactivation n'est souvent cependant que

partielle, une petite partie des transcrits pouvant ne pas être détruite et maintenant une expression faible mais bien présente.

Stratégies de détection des gènes essentiels

De manière générale, la recherche d'essentialité des gènes se base sur l'étude des phénotypes de croissance des mutants obtenus par les techniques d'inactivation précédentes. Les méthodes d'étude des phénotypes de croissance (voir la section précédente) s'appliquent donc également au cas des mutants. Des stratégies particulières ont néanmoins été développées pour augmenter l'efficacité de la recherche des gènes ayant un impact significatif sur la croissance.

Il est tout d'abord utile de distinguer deux manières d'évaluer l'aptitude à croître des mutants, dont les différences ont des conséquences significatives sur l'interprétation de leurs résultats (Gerdes et al. 2006). D'une part, la croissance des mutants peut être évaluée de manière clonale, séparément pour chacun d'entre eux. Le phénotype observé correspond alors directement à l'aptitude brute à croître du mutant. Celle-ci peut également être quantifiée (voir partie précédente) afin de comparer non seulement l'aptitude mais aussi l'efficacité à croître. D'autre part, le second type de test consiste à évaluer l'aptitude à croître des mutants au sein de populations de cellules mélangeant mutants et souches sauvages. Dans cette configuration, la croissance s'effectue en compétition avec les autres souches ; on observe les effets des mutations sur la *valeur sélective* des individus. L'essentialité des gènes est donc définie ici par rapport à leur contribution à l'efficacité de survie de l'organisme (gènes essentiels au succès reproducteur), et non plus seulement par rapport à leur seule capacité à croître (gènes essentiels à la survie). Le choix de la méthode dépend alors de l'exploitation faite des résultats. La première identifie de manière nette les gènes indispensables au phénotype observé, permettant d'investiguer leurs liens, tandis que la seconde, plus large, permet d'identifier des gènes de contributions moindres mais potentiellement importantes du point de vue évolutif.

La stratégie la plus directe de détection des gènes essentiels consiste à inactiver systématiquement chacun des gènes du génome et à tester les phénotypes de croissance des mutants correspondants. Cette approche requiert de pouvoir générer les mutants de manière ciblée. Pour cette raison, les techniques majoritairement utilisées sont les délétions par recombinaison homologue et l'interférence par ARN (Carpenter

& Sabatini 2004), mais des protocoles ont également été développés sur la base de transposons (Kang et al. 2004; Reznikoff & Winterberg 2008; Kobayashi et al. 2003). L'avantage principal de cette stratégie réside dans sa couverture complète du génome, tous les gènes sont systématiquement testés (aux impossibilités expérimentales près). De plus, les mutants créés sont généralement conservés et peuvent être aisément phénotypés ultérieurement pour de nouvelles conditions expérimentales, bénéficiant par exemple alors du débit apporté par des méthodes du type Biolog. E revanche, la création systématique d'un mutant pour chaque gène est une opération lourde, de débit faible.

Afin d'augmenter le débit de l'expérience, des stratégies basées sur l'inactivation aléatoire (ou « shotgun ») des gènes – par transposon (Reznikoff & Winterberg 2008) ou ARN interférent (Ji et al. 2001) – ont été développées. Le principe de ces stratégies consiste à générer un nombre suffisamment élevé de mutants d'inactivation[24] afin que, statistiquement, chacun des gènes ait une probabilité significative d'avoir été inactivé (ou plutôt, une probabilité faible de ne pas avoir été inactivé). En observant ensuite dans les mutants viables à quels endroits sur le génome se retrouvent les transposons (voir Figure 11) ou correspondent les ARN interférents, les gènes non-essentiels peuvent être déterminés. La couverture statistique suffisante des inactivations (appelée *saturation*) permet alors de déduire que les gènes jamais impactés sont essentiels dans les conditions de l'expérience. Comme mentionné plus haut, le test de croissance des mutants peut être réalisé de manière clonale ou au sein d'une population. La méthode de « genetic footprinting », relativement répandue pour rechercher les gènes spécifiquement associés à un phénotype particulier, correspond à ce dernier cas (Smith et al. 1995; Hare et al. 2001). Les lieux d'insertions des transposons (déterminés par PCR[25], voir Figure 11) sont comparés pour deux populations similaires mais cultivées dans des environnements distincts. Les

[24] Dans le cas des transposons, la non spécificité de l'insertion garantit dans une certaine mesure la couverture aléatoire des inactivations. Dans le cas des ARN interférents, des banques aléatoires d'ARN antisens sont généralement créées par fractionnement aléatoire de la séquence génomique (Ji et al. 2001).

[25] PCR : « Polymerase Chain Reaction ». Méthode d'amplification d'une région précise de l'ADN à partir d'oligonucléotides délimitant les extrémités de la région et servant d'amorces à l'ADN polymérase. La région amplifiée est appelée *produit de PCR*.

différences significatives de fréquence d'insertion à certaines localisations du génome révèlent alors l'essentialité conditionnelle des gènes correspondants. Le principal inconvénient des stratégies aléatoires est la faible maîtrise de l'inactivation des gènes, rendant parfois difficile l'interprétation de l'origine de l'essentialité.

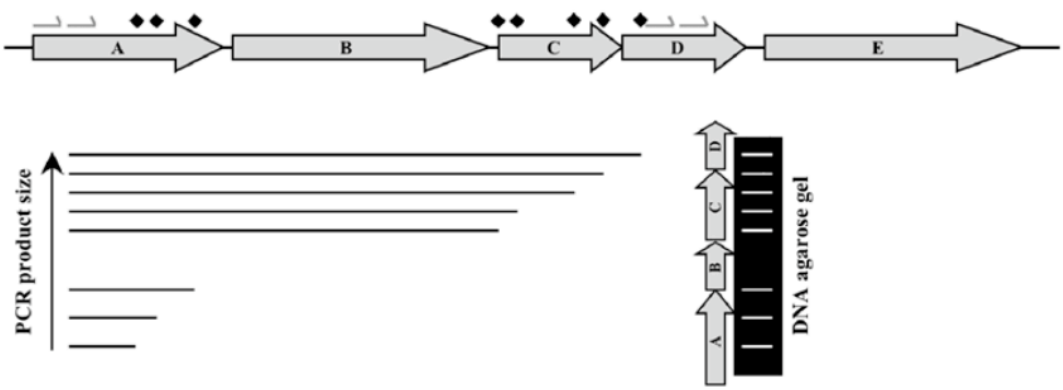


Figure 11. Méthode de « genetic footprinting ». Les lieux d'insertion des transposons sont déterminés par la longueur des produits de PCR entre les amorces choisies à des endroits connus du génome et les amorces placées dans les transposons. Les zones vides du gel d'électrophorèse correspondent aux régions génomiques n'ayant pas retenu d'insertion après sélection des mutants, révélant une possible essentialité des gènes présents à ces loci. Demi-flèches, amorces de PCR ; losanges, lieux d'insertion des transposons. Figure extraite de Scholle & Gerdes (2008).

Ces stratégies furent appliquées à un nombre pour l'instant relativement restreint d'organismes (voir Tableau 3). Cependant, leur accessibilité est en constante amélioration, et il est probable qu'elles occuperont une place plus importante dans la boîte à outils des biologistes moléculaires (Carpenter & Sabatini 2004). S'agissant des résultats d'essentialités existants, il est utile de mentionner les initiatives ayant pour but de les centraliser (Yamazaki et al. 2008; Zhang & Lin 2009).

Organisme	Méthode d'inactivation	Test de croissance des mutants	Référence
A. baylyi ADP1	Délétion ciblée	Clones	(de Berardinis et al. 2008)
M. genitalium, M. pneumonia	Transposon, aléatoire	Population	(Hutchison et al. 1999)
M.genitalium	Transposon, aléatoire	Clones	(Glass et al. 2006)
M. pulmonis	Transposon, aléatoire	Clones	(French et al. 2008)
S. aureus WCUH29	ARN interférent, aléatoire	Clones	(Ji et al. 2001)
S. aureus RN4220	ARN interférent, aléatoire	Clones	(Forsyth et al. 2002)
H. influenzae Rd	Transposon, aléatoire	Population	(Akerley et al. 2002)
S. pneumoniae Rx-1	Disruption ciblée	Clones	(Thanassi et al. 2002)
S. pneumoniae D39	Délétion ciblée	Clones	(Song et al. 2005)
M. tuberculosis H37Rv	Transposon, aléatoire	Population	(Sassetti et al. 2003)
B. subtilis 168	Disruption ciblée	Clones	(Kobayashi et al. 2003)
E. coli K-12 MG1655	Transposon, aléatoire	Population	(Gerdes et al. 2003)
E. coli K-12 MG1655	Transposon, ciblée	Clones	(Kang et al. 2004)
E. coli K-12 MG1655	Délétion ciblée	Clones	(Baba et al. 2006)
P. aeruginosa PAO1	Transposon, aléatoire	Clones	(Jacobs et al. 2003)
P. aeruginosa PA14	Transposon, aléatoire	Clones	(Liberati et al. 2006)
S. typhimurium	Disruption aléatoire	Clones	(Knuth et al. 2004)
H. pylori G27	Transposon, aléatoire	Population	(Salama et al. 2004)
F. novicida	Transposon, aléatoire	Clones	(Gallagher et al. 2007)

Tableau 3. Études expérimentales à grande échelle de l'essentialité des gènes pour des organismes bactériens. Données issues et complétées à partir de Gerdes et al (2006).

2.2.2 Exploitation des données d'essentialité

Historiquement, les premières études d'essentialité de gènes chez les bactéries furent motivées par la recherche de cibles thérapeutiques pour des médicaments anti-infectieux (Ji et al. 2001; Thanassi et al. 2002; Forsyth et al. 2002; Hare et al. 2001; Arigoni et al. 1998; Reich et al. 1999; Chalker & Lunsford 2002). De nombreuses études furent ainsi conduites pour des bactéries pathogènes, notamment dans le cadre des recherches de groupes privés.

Toujours à des fins d'applications pratiques, l'étude des phénotypes d'inactivation de gènes est également utilisée en ingénierie du métabolisme. Elle permet d'identifier des mutations optimisant l'efficacité de production (ou de dégradation, selon l'objectif recherché) de l'organisme utilisé en neutralisant par exemple des voies alternatives en compétition pour les ressources ou des régulations inhibitrices (Park et al. 2008).

Plus fondamentalement, les études portant sur l'évolution des organismes, et notamment de leurs génomes, ont exploité avec intérêt les résultats d'essentialité des gènes. Un grand nombre d'entre elles ont ainsi cherché à établir des corrélations entre l'essentialité des gènes et leurs caractéristiques évolutives, par exemple la vitesse d'évolution ou les biais de leurs séquences, leur conservation entre les espèces ou leur position dans le génome (Fang et al. 2005; Gong et al. 2008; Papp et al. 2004; Harrison et al. 2007; Rocha & Danchin 2003). Ces analyses sont motivées par l'exploration des mécanismes d'évolution ; la distinction entre gènes essentiels et gènes non-essentiels permet d'une part d'estimer l'impact de la valeur sélective des gènes sur leur évolution et d'autre part d'évaluer l'importance de la robustesse aux perturbations génétiques comme caractère marquant de l'évolution. Une autre partie des études liées à l'évolution se sont basées sur l'hypothèse que les gènes essentiels représentent des fonctions universellement requises, devant donc être retrouvées dans chaque organisme. En combinant données d'essentialité et analyses de la conservation des gènes entre organismes, ces études ont cherché à élucider des scénarios évolutifs et à reconstruire des génomes ancestraux (Koonin 2003). De manière connexe, de nombreuses initiatives ont été entreprises pour construire des génomes minimaux, à la fois via des méthodes bioinformatiques ou expérimentales (Koonin 2003; Glass et al. 2006; Mushegian & Koonin 1996).

Enfin, et plus proche du sujet de cette thèse, les phénotypes de croissance de mutants sont aussi largement utilisés pour rechercher les fonctions de gènes et comprendre le fonctionnement de processus biologiques. Ces approches sont basées sur la recherche de liens de causalités spécifiques entre la présence d'un gène et l'occurrence d'un phénotype, afin de guider soit la recherche des gènes impliqués dans la réalisation d'une fonction particulière (approche de *génétique classique*), soit la recherche de fonctions biologiques associées à un gène particulier (approche de

génétique inverse, voir Figure 12). À ces deux types d'approches correspondaient traditionnellement des types d'expériences distinctes, par exemple le « genetic footprinting » en génétique classique ou le phénotypage à haut débit en génétique inverse. La réalisation de banques de mutants d'inactivation à grande échelle permet désormais de lier les deux approches, où les phénotypes de chacun des mutants peuvent être systématiquement testés (Carpenter & Sabatini 2004). Ces approches sont utilisées à des fins exploratoires (Aghaie et al. 2008) mais également de confirmation ou d'invalidation de fonctions de gènes, lorsque celles-ci sont attribuées sur la base d'indices faibles (de Berardinis et al. 2008; Joyce et al. 2006; Baba et al. 2006). Les processus métaboliques se prêtent bien à l'utilisation de ces approches (Gerdes et al. 2006), qui ont d'ailleurs largement contribué à l'identification des gènes impliqués dans les voies métaboliques connues. En effet, des tests phénotypiques caractérisant assez précisément une fonction métabolique peuvent être élaborés en combinant complémentation par des substrats et inactivation de voies métaboliques. Une formalisation de cette démarche a d'ailleurs été récemment développée et mise en pratique dans un robot réalisant automatiquement à la fois les raisonnements et les expériences correspondant à ce type d'approche (King et al. 2009; King et al. 2004).

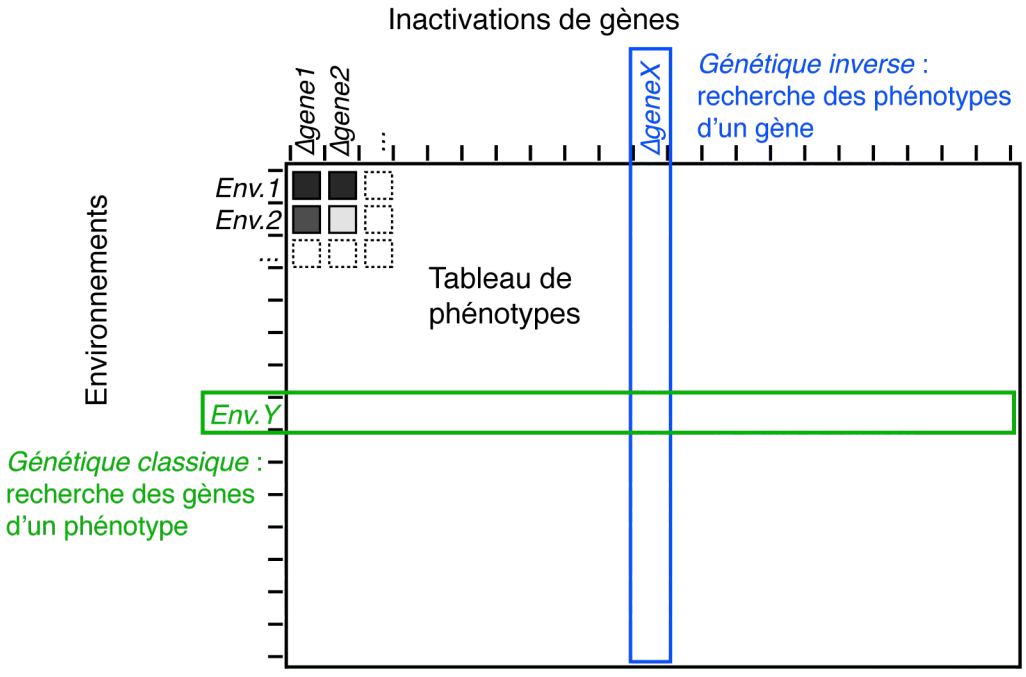


Figure 12. Principes des approches de génétique classique et de génétique inverse.

Toutes ces études reconnaissent cependant l'existence de difficultés dans l'utilisation des données d'essentialités de gènes. Tout d'abord, comme souligné plus haut, l'effet phénotypique d'une inactivation de gène s'interprète parfois de manière

ambiguë. La non-viabilité d'un mutant dépend en effet fortement des conditions de croissance de celui-ci et de sa mise en compétition ou non au sein d'une population de cellules. L'environnement de croissance a de plus un effet majeur sur l'essentialité des gènes, surtout pour ceux jouant un rôle dans le métabolisme. Cet aspect est mis à profit pour justement identifier les gènes spécifiques à un environnement dans un organisme donné, mais il brouille les comparaisons d'essentialité entre organismes (Gerdes et al. 2006). Enfin, et surtout, l'essentialité d'une fonction biologique dans un organisme n'implique pas nécessairement l'essentialité du ou des gènes associés à sa réalisation. La présence de mécanismes alternatifs ou de gènes de fonctions redondantes peut en effet rendre chacun des gènes individuellement non-essentiels. Cette robustesse aux perturbations génétiques motive par ailleurs de nombreuses études (Papp et al. 2004; Kuepfer et al. 2005; Deutscher et al. 2006; Stelling et al. 2004; Kitano 2007) mais rend la recherche de fonctions essentielles plus délicate. Le recours aux délétions multiples permet d'aller un cran plus loin dans cette recherche (Tong et al. 2004; Butland et al. 2008), mais l'explosion du nombre de combinaisons à tester rend impossible l'utilisation naïve de ces approches.

Dans le cas du métabolisme, la connaissance des voies métaboliques et des associations entre gènes et activités réactionnelles peut aider à interpréter correctement les résultats d'essentialité au regard de l'environnement et des potentielles redondances (Gerdes et al. 2006). Cependant, comme déjà mentionné plus haut, la complexité du métabolisme et le grand nombre de résultats à interpréter rendent ces analyses souvent difficiles à réaliser. La modélisation du métabolisme peut justement assister l'investigateur dans cette tâche en réalisant ces raisonnements automatiquement. Ceci constitue le sujet des travaux de notre thèse qui s'inscrit dans un mouvement global d'initiatives en ce sens, dont nous effectuerons une revue dans la partie suivante.

3 Modélisation du métabolisme

Dans cette section, nous donnerons tout d'abord une rapide vue d'ensemble des méthodes de modélisation du métabolisme – avec un point de vue orienté vers la prise en compte de l'ensemble du métabolisme de la cellule – avant de présenter plus en détail la méthode de modélisation retenue dans cette thèse. En dernier lieu, nous

effectuerons un état de l'art à la date du début de la thèse – fin 2005 – sur l'exploitation des phénotypes de croissance et des essentialités de gènes à l'aide des modèles métaboliques.

3.1 Approches de modélisation du métabolisme

Largement employées en physique, mécanique ou chimie, la modélisation et la simulation informatique ne sont en comparaison utilisées que depuis récemment en biologie, à l'exception notable de l'écologie et de l'épidémiologie dans lesquelles les mathématiques occupent une place significative depuis longtemps (May 2004). Les systèmes physico-chimiques étudiés en biologie ont en effet longtemps été jugés difficiles à aborder par ces approches du fait de leur grande complexité et surtout de la part importante d'inconnu dans leur fonctionnement. Cependant, depuis quelques décennies, des progrès considérables ont été effectués dans leur compréhension grâce notamment aux avancées technologiques qui permettent de caractériser un nombre toujours croissant de leurs acteurs et interactions. La reconnaissance toujours présente (et même accrue) de la complexité de ces systèmes associée à la disponibilité d'informations sur leurs acteurs ont alors motivé le développement d'approches plus formelles pour la compréhension globale de ces systèmes[26] dans lesquelles les mathématiques et la modélisation jouent un rôle primordial. Le projet Physiome, consacré à l'étude du fonctionnement du cœur par la modélisation à différentes échelles – moléculaire, cellulaire, de l'organe entier – et de différentes composantes – mécanique, biochimique, électrique –, est un exemple phare de ce type d'approche (Noble 2002; Hunter & Borg 2003).

S'agissant du métabolisme, une variété relativement large de méthodes de modélisation ont été élaborées, dont la nature dépend souvent à la fois des questions posées et de la « culture » scientifique – informatique, mathématique, physique – de leurs auteurs. En première approximation, on peut distinguer ces méthodes selon le niveau de détail de leurs prédictions (Figure 13) (Stelling 2004).

[26] Désignées communément sous le terme de biologie des systèmes (Kitano 2002; Stelling 2004).

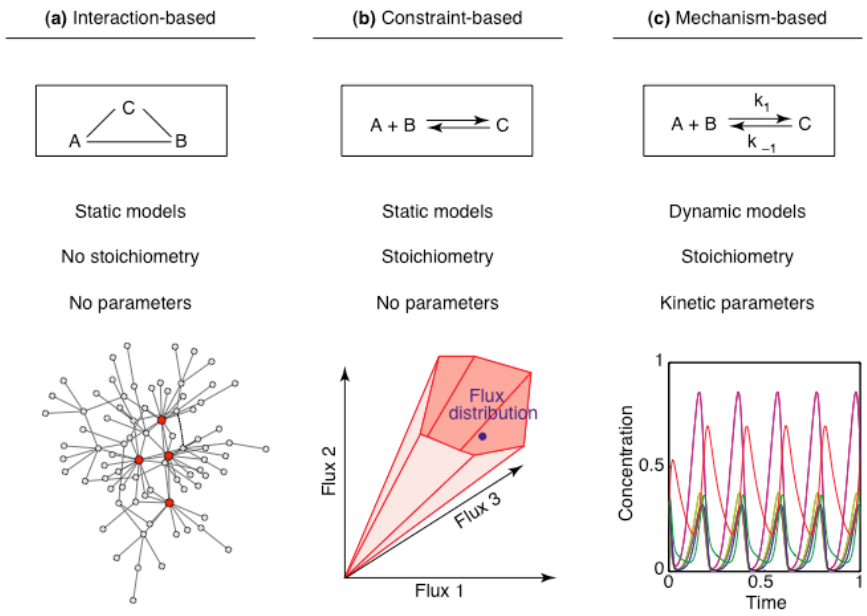


Figure 13. Formalismes de modélisation du métabolisme classés selon leurs niveaux de détails. (a) modélisation sous forme de graphes, construits à partir d'informations sur les interactions entre les acteurs du métabolisme, (b) modélisation à base de contraintes, tenant compte des relations quantitative entre flux de réaction lorsque le métabolisme opère en régime stationnaire, (c) modélisation cinétique, représentant l'évolution temporelle des quantités d'enzymes et de métabolites. Les illustrations sur la ligne inférieure représentent des résultats typiques de ces méthodes : (a) métabolites centraux (liés à un grand nombre de métabolites) en rouge dans un réseau métabolique, (b) ensemble des valeurs de flux réactionnels possibles en régime stationnaire (pour plus de détails su ce formalisme, voir section 3.2), (c) dynamique de la concentration de métabolites. Figure issue de Stelling (2004).

La méthode de modélisation a priori la plus naturelle pour un physicien consiste à décrire l'évolution dans le temps des quantités de métabolites et d'enzymes ; il s'agit de la *modélisation cinétique* ou *modélisation quantitative* (Di Ventura et al. 2006). Pour cela, des modèles mathématiques de la cinétique des enzymes (voir la section 1.2.4 et Cornish-Bowden (2004)) sont utilisés pour représenter les dépendances entre ces grandeurs, le plus souvent sous la forme d'équations différentielles. Leur résolution analytique est le plus souvent impossible sauf dans les cas très simples. On utilise alors la simulation informatique pour obtenir une solution numérique – des courbes d'évolution dans le temps des grandeurs – ou des outils théoriques, tels que les analyses de bifurcation et de stabilité, pour caractériser le comportement du système (Di Ventura et al. 2006). La complexité des modèles cinétiques varie significativement, en fonction des hypothèses simplificatrices formulées et de la taille du réseau métabolique considéré. Ainsi, certains modèles prennent en compte la localisation spatiale des molécules dans la cellule (Lemerle et al. 2005; Moraru et al. 2008) tandis que d'autres (la majorité) supposent une répartition homogène. De même, la sensibilité des processus aux fluctuations aléatoires peut nécessiter un

traitement stochastique de ces phénomènes, rendant leur résolution plus complexe (Di Ventura et al. 2006; Gillespie 2007). Les modèles cinétiques ont été employés avec succès pour analyser le comportement dynamique de petits systèmes métaboliques et prédire leurs réponses à des perturbations (Klipp et al. 2002; Zaslaver et al. 2004). Leurs applications les plus significatives jusqu'à présent portent toutefois plutôt sur les processus de signalisation ou de régulation transcriptionnelle (Di Ventura et al. 2006; Barkai & Leibler 1997; Bonneau et al. 2007).

Les phénotypes de croissance sont une manifestation globale du fonctionnement du réseau métabolique. Leur étude requiert donc de tenir compte de l'ensemble des réactions. La modélisation cinétique ne peut satisfaire actuellement cette contrainte. D'une part, les comportements cinétiques des enzymes ne sont de loin pas tous caractérisés[27] et reposent sur de nombreux paramètres numériques souvent inconnus[28]. D'autre part, le nombre de réactions impliquées rend les analyses extrêmement complexes et souvent difficiles à réaliser, même par simulation numérique. Pour ces raisons, des cadres de modélisation de moindres capacités prédictives – prédictions moins précises ou hypothèses plus restrictives – mais aux formalismes utilisables à des tailles de réseaux métaboliques plus importantes ont été élaborés.

L'*analyse du contrôle métabolique* a pour objectif de quantifier les dépendances entre les différentes grandeurs d'un système métabolique – flux de réaction, concentrations de métabolites ou d'enzymes – fonctionnant en première approximation autour d'un régime stationnaire (Kacser & Burns 1973; Heinrich & Rapoport 1974; Fell 1992). Ce type d'analyse détermine typiquement des *coefficients de contrôle* exprimant dans quelle mesure les variations de certaines grandeurs influent sur les autres grandeurs et le comportement global du système. L'application de ce type d'analyse à des voies métaboliques linéaires a par exemple pu montrer que le contrôle du flux à l'état stationnaire d'une voie métabolique de ce type se répartit entre les différentes enzymes de cette voie et n'est pas simplement déterminé par une

[27] Même si des initiatives cherchent à établir des cinétiques « génériques » pour les enzymes (Liebermeister & Klipp 2006).
[28] Malgré l'existence de bases de données centralisant les informations sur ces paramètres (Barthelmes et al. 2007; Wittig et al. 2006).

seule d'entre elle[29], « l'étape limitante » (Fell 1992). D'un point de vue plus global et plus proche des phénotypes de croissance, ce type d'analyse a également été utilisé pour étudier les dépendances entre l'efficacité à se reproduire d'organismes et les flux dans certaines de leurs voies métaboliques clés (Dykhuizen et al. 1987). Néanmoins, ces analyses nécessitent toujours de déterminer un nombre relativement élevé de paramètres numériques (quantifiant les dépendances), bien que plus réduit que ceux des modèles cinétiques. Pour cette raison, l'analyse du contrôle métabolique est majoritairement appliquée dans deux cas de figures distincts : (1) la démonstration théorique d'un type de comportement métabolique et (2) l'étude précise du comportement d'un ensemble de quelques voies métaboliques en exploitant des données expérimentales.

À un niveau de simplification supplémentaire se situe la *modélisation à base de contraintes* (Price et al. 2004), que nous avons adoptée dans cette thèse et dont nous présenterons le formalisme et les références majeures dans la section suivante (section 3.2). Ce cadre de modélisation se concentre exclusivement sur l'étude des régimes stationnaires du métabolisme, mais sans chercher à quantifier leurs dépendances aux variations des grandeurs du système tel que le fait l'analyse du contrôle métabolique. Dans un souci de simplicité, il décrit le fonctionnement du métabolisme uniquement avec les flux de réaction. Plutôt que de chercher à déterminer la valeur précise de ces flux, tâche difficile et nécessitant une grande quantité d'information (c'est un des objectifs des modèles cinétiques), le principe de ces modèles consiste au contraire à exploiter au mieux l'information disponible pour affiner progressivement la caractérisation des flux métaboliques. Ces modèles raisonnent ainsi sur des ensembles de valeurs de flux possibles compte tenu de l'information disponible : peu d'information résulte en de grands ensembles de valeurs possibles tandis que l'ajout d'information réduit leurs tailles (et affine donc la connaissance des flux). L'information est prise en compte dans ces modèles sous forme de contraintes mathématiques sur les flux. Celles-ci peuvent simplement définir des plages de valeurs connues (ou mesurées) pour certaines réactions ou traduire des dépendances complexes entre flux. L'hypothèse de régime stationnaire entre dans ce dernier cas ;

[29] La répartition du contrôle n'est cependant pas homogène et, quand bien même il n'existe pas une unique étape limitante, le but de l'analyse du contrôle métabolique est de déterminer lesquelles contribuent le plus significativement au contrôle.

elle se traduit mathématiquement par des relations linéaires entre flux exprimant la conservation de la matière. Le principal atout de cette méthode est donc de pouvoir gérer le manque d'information et de pouvoir ainsi être appliquée pour des systèmes de tailles plus conséquentes que pour les modèles cinétiques. Dans la pratique, son utilisation pour des réseaux métaboliques globaux permet d'obtenir des prédictions intéressantes sur la valeur de leurs flux, notamment grâce au fait que la contrainte de régime stationnaire puisse être appliquée à cette échelle[30] et contribue à affiner significativement la caractérisation des flux. Nous reviendrons plus en détail sur ce cadre de modélisation dans la partie suivante.

La représentation du réseau métabolique sous forme de *graphe* permet d'en simplifier encore plus sa modélisation (voir Figure 13). Un graphe est un concept mathématique et informatique permettant de représenter des liens (éventuellement orientés) entre objets ; il se compose simplement d'un ensemble d'objets et d'un ensemble de liens entre objets[31]. Les développements de la théorie des graphes ont apporté un vaste panel de méthodes pour explorer leurs propriétés : recherche de chemins entre objets à travers les liens du graphe, statistiques topologiques, recherche de motifs topologiques caractéristiques, décomposition en sous-graphes de densités de liens plus élevées… De par sa nature, le réseau métabolique se prête bien à l'utilisation des graphes. Il peut être représenté sous la forme d'un graphe simple où les objets sont les réactions ou les métabolites et les liens indiquent que les réactions (respectivement les métabolites) partagent un ou plusieurs métabolites (respectivement une ou plusieurs réactions). Il peut être également représenté de manière plus complète en utilisant un graphe à deux types d'objets[32] dans lequel à la fois les métabolites et les réactions sont représenté ; les liens associent alors les métabolites aux réactions auxquelles ils participent (voir Figure 14).

[30] La seule information requise est la stœchiométrie des réactions, qui est en général connue dans le métabolisme.
[31] La nomenclature usuelle appelle les objets *nœuds* et les liens *arêtes*.
[32] Graphe *biparti*.

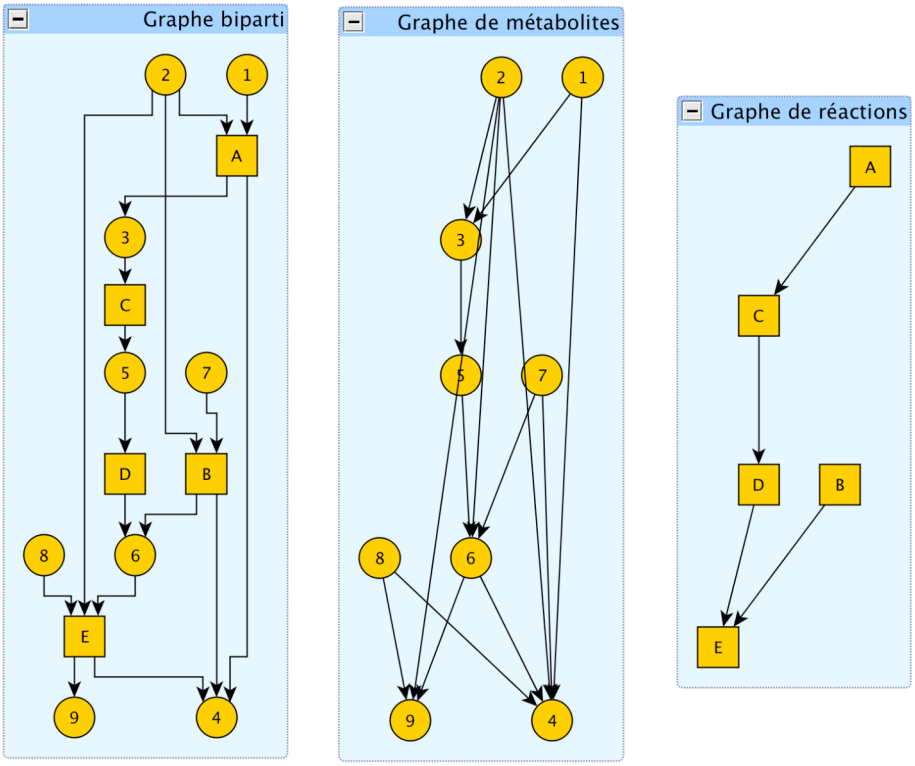


Figure 14. Représentations sous forme de graphes d'un réseau métabolique théorique. Les métabolites sont représentés par des cercles et désignés par des nombres, les réactions représentées par des carrés et désignées par des lettres. À gauche, représentation sous forme d'un graphe biparti ; au centre, graphe simple de métabolites ; à droite, graphe simple de réactions.

La simplicité des graphes métaboliques permet leur utilisation dès lors que les réactions du réseau métabolique sont connues. C'est pourquoi ils ont connu un intérêt prononcé au moment où les réseaux métaboliques globaux de plusieurs organismes ont été reconstruits, à la suite du séquençage et de l'annotation de leurs génomes. Une première catégorie d'études s'est principalement attachée à analyser la structure de ces graphes, dans le but de mettre en évidence des propriétés structurelles communes entre organismes (Jeong et al. 2000) ou de décomposer les réseaux en modules fonctionnels similaires à la notion de voie métabolique (Ravasz et al. 2002). Une seconde catégorie d'études a quant à elle été consacrée à élaborer des algorithmes permettant d'explorer les capacités de conversion des réseaux métaboliques en fonction de leurs environnements. Ces études – basées sur des méthodes dites *d'expansion de réseau* – permettent typiquement de générer l'ensemble des métabolites pouvant être potentiellement synthétisés par un réseau de réactions à partir d'un ensemble initial de métabolites (Handorf et al. 2005; Romero & Karp 2001; Raymond & Segrè 2006). La simplicité extrême des graphes métaboliques limite cependant leurs capacités prédictives. Les aspects quantitatifs, et notamment la

stœchiométrie des réactions, sont en effet ignorés malgré leur importance dans la compréhension du fonctionnement métabolique (de Figueiredo et al. 2009). Ils sont donc majoritairement utilisés lorsque la reconstruction des réseaux métaboliques ne peut être effectuée que de manière grossière – notamment pour les études comparatives de nombreux réseaux – ou lorsque la taille des réseaux nécessite une modélisation « allégée ».

Des initiatives ont cependant cherché à améliorer les capacités prédictives des graphes métaboliques en les étendant au sein de cadres de modélisation informatique[33] plus élaborés (Fisher & Henzinger 2007). Parmi ces derniers, les réseaux de Petri se sont révélés être particulièrement bien adaptés à l'étude du métabolisme, permettant notamment d'aborder de façon qualitative la dynamique de certaines voies métaboliques (Simão et al. 2005; Hofestädt 2003; Reddy et al. 1996; Koch et al. 2005). Ces cadres de modélisation passent toutefois encore difficilement à l'échelle du métabolisme cellulaire tout entier, pour lequel prédomine l'utilisation des graphes ou des modèles à base de contraintes.

3.2 Les modèles à base de contraintes : reconstruction et applications

Cette partie présente de manière détaillée la modélisation à base de contraintes. Elle couvre trois aspects : (1) le formalisme mathématique sous-jacent, (2) la reconstruction pratique de ces modèles, notamment les méthodes et ressources bioinformatiques de reconstruction des réseaux métaboliques (évoquées dans la section 1.4), et (3) ses principales applications. Dans un premier temps, nous invitons le lecteur à lire un article de revue – que nous avons rédigé au cours de la thèse pour le journal *FEMS Microbiology Reviews* (Durot et al. 2009) – traitant des deux

[33] Un modèle informatique se distingue d'un modèle mathématique de par son exécution directe par un ordinateur (Fisher & Henzinger 2007). Les modèles mathématiques sont généralement exprimés par des équations dont la résolution n'est pas nécessairement simple à réaliser. L'informatique peut aider à leur résolution en utilisant des programmes dédiés (notamment la simulation numérique). A l'inverse, les modèles informatiques sont exprimés sous la forme d'un langage ou d'un algorithme pouvant être directement exécuté par l'ordinateur. Ils se basent sur la description d'états et spécifient sous quelles conditions ces états évoluent. Ils sont donc par construction de nature qualitative.

derniers aspects. Nous donnerons dans un deuxième temps des précisions au lecteur sur le cadre mathématique de la modélisation.

3.2.1 Article de revue

REVIEW ARTICLE

Genome-scale models of bacterial metabolism: reconstruction and applications

Maxime Durot, Pierre-Yves Bourguignon & Vincent Schachter

Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, Evry, France

Correspondence: Vincent Schachter, Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, CP5706, 91057 Evry, Cedex, France. Tel.: +33 1 60 87 25 92; fax: +33 1 60 87 25 14; e-mail: vs@genoscope.cns.fr

Received 30 July 2008; revised 22 October 2008; accepted 22 October 2008.
First published online December 2008.

DOI:10.1111/j.1574-6976.2008.00146.x

Editor: Victor de Lorenzo

Keywords
metabolic network; systems biology; computational methods; genome-scale metabolic models; metabolic engineering; omics data integration.

Abstract

Genome-scale metabolic models bridge the gap between genome-derived biochemical information and metabolic phenotypes in a principled manner, providing a solid interpretative framework for experimental data related to metabolic states, and enabling simple *in silico* experiments with whole-cell metabolism. Models have been reconstructed for almost 20 bacterial species, so far mainly through expert curation efforts integrating information from the literature with genome annotation. A wide variety of computational methods exploiting metabolic models have been developed and applied to bacteria, yielding valuable insights into bacterial metabolism and evolution, and providing a sound basis for computer-assisted design in metabolic engineering. Recent advances in computational systems biology and high-throughput experimental technologies pave the way for the systematic reconstruction of metabolic models from genomes of new species, and a corresponding expansion of the scope of their applications. In this review, we provide an introduction to the key ideas of metabolic modeling, survey the methods, and resources that enable model reconstruction and refinement, and chart applications to the investigation of global properties of metabolic systems, the interpretation of experimental results, and the re-engineering of their biochemical capabilities.

Introduction

The flow of genome sequencing, metagenome sequencing and other high-throughput experimental efforts aimed at exploring the space of microbial biochemical capabilities has been steadily growing in recent years. At the time of writing, more than 1800 bacterial genome-sequencing projects have been initiated and nearly 650 have been completed (http://www.genomesonline.org, http://www.ebi.ac.uk/integr8). Combined with increasingly efficient annotation methods, these set the stage for the systematic identification of most enzymes encoded in the genomes of the corresponding bacterial species. A variety of so-called '-omics' technologies now routinely provide large-scale functional clues on molecular interactions and cellular states, offering snapshots of the dynamic operation of metabolism under specified conditions, and adding to the store of accumulated knowledge on microbial biochemistry and physiology.

Simultaneously, the expected wealth of new biochemical activities, the progress of metabolic engineering techniques aimed at harnessing these activities, and the perspective of applications to white and green biotechnology have triggered a strong renewed interest in the exploration of bacterial metabolism. In addition to charting the range of naturally evolved chemical transformations, relevant research questions include the following: How does the global metabolism of a bacterium react to changes in its environment? What kind of joint metabolic operation of distinct species can help sustain a bacterial community? How can genomic and biochemical information be best exploited to gain insights into the relationship between an organism's genotype and its phenotype? For instance, can we predict changes in metabolism-related phenotypic traits caused by simple or complex genotype modifications? How did metabolic processes evolve? How can metabolic networks be efficiently reprogrammed for a variety of utilitarian purposes?

Investigations of a bacterium's metabolism are typically fed by knowledge (ultimately from observations) at two different scales of description of the chemistry at work within cells. The larger scale focuses on the physiology of the whole bacterial

cell. For instance, which media is it able to grow on? What are the relative quantities of chemical nutrients it requires for growth? How efficient is the cell at converting chemicals from the environment into its own components? Such metabolic capabilities result from the coordinated action of the enzymes expressed in the respective species, the knowledge of which belongs to the finer, molecular scale. Each of the corresponding biochemical conversions can be identified either directly by performing enzymatic assays, or indirectly, from the genome sequence, through a homology relationship with proteins whose function has been previously elucidated. Together, the reactions that have been demonstrated to potentially occur in the cell form the *metabolic network* of the organism. Metabolic networks can thus be viewed as lists of those molecular mechanisms (reactions) and associated molecular components (enzymes, substrates, and products) that are most directly related to the metabolic capabilities mentioned above.

For a given bacterial species, confronting knowledge from these two scales, molecular vs. cellular, can reveal inconsistencies. For instance, it may happen that no sequence of identified reactions is capable of producing one of the essential cell components from the set of compounds available in a defined growth medium, even though the species is known to grow on that medium. Furthermore, when the two scales are consistent, their relationship can be investigated further in order to enumerate the possible implementations of the physiology that the metabolic network can achieve. Biochemists have traditionally performed such investigations by modularizing the set of reactions into *metabolic pathways*, typically grouping together reactions that allow the conversion of one or more 'input' metabolites into 'output' metabolites. Pathways boundaries are somewhat arbitrary, even though inputs and outputs tend to be metabolites involved in several reactions. Pathway-based analyses are thus focused on the possible fates of a restricted number of compounds, and are amenable to manual expertise thanks to the simplification brought by the modularized view (Huang*et al.*, 1999; Teusink *et al.*, 2005; Risso *et al.*, 2008).

Yet, metabolic pathways typically involve a large number of 'side metabolites' such as cofactors and byproducts of chemical reactions, and metabolism is as much about converting nutrient into cell components as it is about regenerating cofactors and recycling (or secreting) ultimately unused byproducts. The latter transformations typically involve several pathways, and are dependent on the stoichiometry and rates of the reactions. Manual approaches are insufficient to assess their feasibility by a given network for at least two reasons: metabolic networks are too large, and the question requires a quantitative analysis.

Bridging that gap between knowledge of the metabolic network structure and observed metabolic phenotypes is precisely where metabolic models come into play. Generally speaking, a model of a natural system is one of many possible mathematical representation of that system, explicitly describing some of its features and supporting predictions on some other features, the latter being typically time- or environment dependent. In this particular case, knowledge of the metabolic network alone is not quite sufficient to predict the metabolic capabilities of a cell. Also needed are a structured (mathematical) representation of that network, together with a set of rules and possibly quantitative parameters enabling simulations or predictions on the joint operation of all network reactions in a given environment, and in particular predictions on the values of metabolite fluxes and/or concentrations (Papin *et al.*, 2003). The above, in short, constitutes a metabolic model.

Constraint-based genome-scale models of metabolism (Palsson, 2006) are a category of models precisely aimed at assessing the physiological states achievable by a given metabolic network, and at uncovering their biochemical implementation in terms of metabolic fluxes. They offer an idealized view of the cell as a set of 'pipes,' with metabolites flowing through each pipe, and biochemical conversions taking place at junctions between pipes. Some metabolites can also be exchanged with the environment, flowing in or out of the system through dedicated pipes that can be opened or shut, and may have upper bounds on their throughput. The cell is required to achieve balanced production and consumption of all the intermediate substrates and products involved in its metabolism: what flows in a junction must flow out.

Constraint-based models can help investigate in a systematic manner most of the research questions listed at the start of this introduction, because they provide a way to explore the consequences on the operation of the entire metabolic network of the piecemeal information available on each of its parts. They are especially well suited to 'what if' experiments involving genetic or environmental perturbations, such as: how would the cell behave in an environment with a different chemistry than the ones that have been experimented on? How would one or more deletions affect its metabolic capabilities? Which deletions would maximize the production of both metabolite x and biomass?

Before a model for a given species can be used to gain new insights into its metabolic capabilities or evolutionary history, it must first be built from the scattered genomic, biochemical, and physiological information available on that species up to a point where known physiology can be predicted from biochemistry without major mistakes. This process is sometimes known as 'model reconstruction'; its endpoint is a functional genome-scale model, i.e. a structured representation of the current state of knowledge on the metabolism of the respective species (Reed *et al.*, 2006a). The model provides a framework to interpret new experimental data gathered at the cellular or molecular scale. That data may be incompatible with the current model, in

which case either or both should be questioned, leading to possible revisions or improvements. If, on the other hand, data and model are compatible, the new evidence may still narrow down the set of possible metabolic behaviors of the cell, thus enriching the model (Covert *et al.*, 2004).

This review article covers both the reconstruction of genome-scale metabolic models and their applications to basic and applied research in microbiology. Following a primer on constraint-based models, we will review the state of the art in model reconstruction. Next, we will survey the main applications of metabolic models, from phenotype predictions to data interpretation or metabolic engineering. Practical aspects of direct relevance to the working microbiologist will be covered by a sketch of the main dedicated database and software resources. We will conclude the review with a discussion on future directions in the field.

Foundations of genome-scale metabolic modeling

The metabolic state of a cell and its variation over time can be described by metabolite concentrations and reaction rates, which can be viewed as the 'endpoints' of metabolic operation. These quantities are related by the law of conservation of matter, which states that the net production rate of a metabolite equals the sum of the rates of the reactions consuming or producing it, weighted by the associated relative stoichiometric coefficients. Conversely, enzyme kinetics express reaction rates as complex functions of metabolite concentrations and enzymatic activities, which vary over time as a result of transcriptional and metabolic regulation (Smallbone *et al.*, 2007). Deriving meaningful predictions from these two types of equations for large metabolic systems is a very challenging proposition, not only because of the mathematics, but also because many of the parameters are not known, difficult to measure, and possibly context dependent. In practice, these pitfalls restrict the use of kinetic modeling to metabolic systems much smaller than 'whole-cell' metabolic networks, which typically include hundreds of reactions for a bacterium.

Constraint-based models bypass these difficulties by focusing on the average reaction rates achievable by cells grown in steady or slowly varying environmental conditions. Rates are typically averaged over minutes, fitting with the typical time scale of uptake or secretion rates measurements. Such averages are not affected by transient states because the characteristic relaxation time of metabolic systems – i.e. the time it takes for chemical reactions within the cell to reach a steady state – is much shorter than a minute. Moreover, because environmental changes and variations of enzyme concentrations occur on longer time scales, one need not take into account regulatory changes to assess average reaction rates over minutes. Turnover rates of most intracellular metabolites are high in bacterial cells (Stephanopoulos *et al.*, 1998). At the time scale considered here, their concentrations have therefore generally reached steady levels, and remain constant as long as environmental conditions do not change. As a consequence, the law of conservation of matter constrains the production and consumption rates of these metabolites to be balanced. These assumptions are usually summarized under the expression *steady-state hypothesis* and the corresponding constraint on reaction rates as a *mass balance* (or stoichiometric) *constraint* (Stephanopoulos *et al.*, 1998). Obviously, this reasoning applies only to metabolites that are neither taken in from an external pool (e.g. nutrients) nor excreted from the cell or accumulated in large quantities (e.g. cell components such as nucleic acids, amino acids, or some lipids). For each metabolite that can be 'balanced,' the mass balance constraint can be expressed mathematically by a linear equation relating reaction rates of the form $\sum s_j v_j = 0$, where s_j is the stoichiometric coefficient of the metabolite in reaction j, and v_j the rate of reaction j.

In addition to mass balance constraints, reactions that are known to be thermodynamically irreversible *in vivo* are constrained to have a non-negative reaction rate. Similarly, upper bounds on the reaction rates can be known from measurements or theory and included in the model as additional constraints on the reaction fluxes (Reed & Palsson, 2003).

Mass balance, irreversibility and upper-bound constraints result from the application of simple laws of physics to individual reactions or metabolites from the network. These constraints propagate from reaction to reaction throughout the metabolic network; the constraint-based modeling framework is designed to automatically compute the resulting balance. To that end, it makes use of a succinct mathematical representation of all reaction stoichiometries: the *stoichiometric matrix* (see Fig. 1). In this matrix, columns represent reactions and rows metabolites. The stoichiometric coefficient of a metabolite within a reaction is included at the intersection of the corresponding row and column (see Fig. 1). Reaction rates are represented in constraint-based models by single numbers, the *reaction fluxes*, which are normalized by the weight of the cells harboring the reactions to account for the size of the colony (a reaction flux is typically expressed with the Unit mmol h^{-1} g^{-1} dry wt). Because the goal is to describe the joint operation of many metabolic reactions, it is convenient to define a *flux distribution* as a collection of reaction fluxes covering the entire system. Under the steady-state approximation, the concentrations of balanced metabolites being constant, a flux distribution carries sufficient information to completely describe a state of the system. Using the stoichiometric matrix, a simple matrix equation – summarizing all mass balance equations shown above – can then be used to

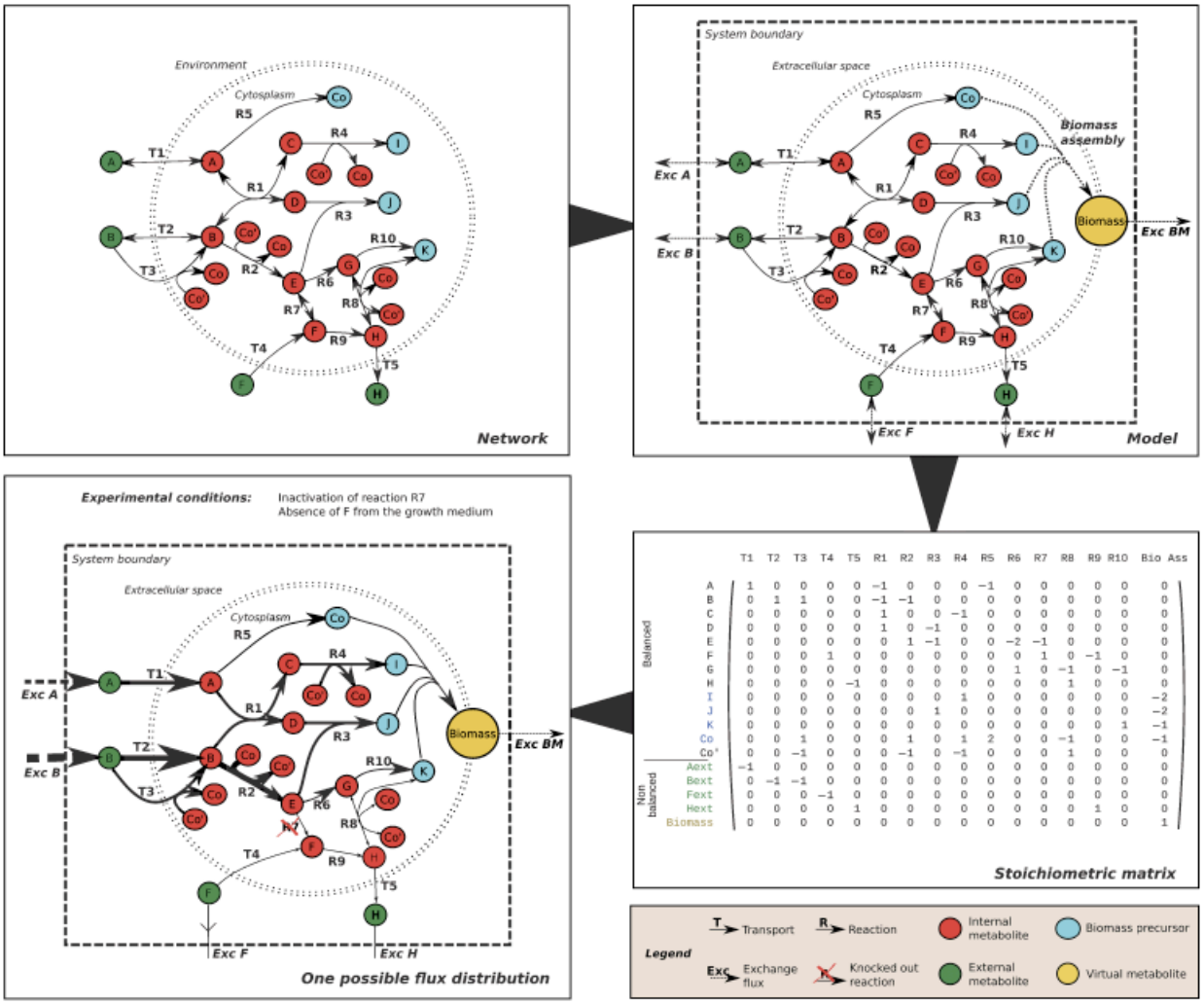


Fig. 1. Genome-scale modeling of metabolism. A metabolic network (top left) is transformed into a model by defining the boundaries of the system, a biomass assembly reaction, and exchange fluxes with the environment (top right). Using the corresponding stoichiometric matrix (bottom right), the achievable flux distributions compatible with enforced constraints can be found (a particular one is depicted in the bottom left figure).

enforce the mass balance constraints on all reactions fluxes: $S.v = 0$, where S is the stoichiometric matrix and v the flux distribution represented as a vector.

A precise definition of the boundary of the system to be modeled is also needed to formulate an explicit mathematical representation. The system typically includes the whole cell and its vicinity, in order to encompass all the exchanges of matter between the cell and its environment. Transport reactions that allow for exchange of specific metabolites with the extracellular space through the membrane are also included in the model. Environmental conditions are then modeled by acting on the balance of the external metabolites: metabolites that are available from the environment can be taken up by transporters while the others can only be excreted.

A flux distribution that is compatible with all the constraints in a given environment is considered achievable (or feasible) by the cell, whereas a distribution that violates at least one of these constraints is not. The simplicity of the system of linear equations that represent constraints is one of the main strengths of the framework, because it permits fast assessments of the feasibility of a flux distribution using a computer and standard algorithms.

The simplicity of constraint-based models comes at the expense of a number of limitations in their predictive capabilities. Such models focus solely on reaction fluxes, and completely ignore the influence of metabolites and enzymes. In reality, however, enzyme kinetics, and transcriptional or metabolic regulation may significantly influence reaction fluxes. Regulation can for instance limit the use of a pathway by downregulating some of its enzymes when particular environmental conditions are met. These mechanisms, if they could somehow be taken into account,

would eliminate flux distributions otherwise allowed by constraint-based models. In other words, models may allow 'false-positive' metabolic states, which respect the enforced metabolic constraints but are inconsistent with other biological processes. Several attempts have been made to extend the constraint-based modeling framework, in order to account for regulatory interactions (Covert *et al.*, 2001), signaling processes (Lee *et al.*, 2008b), the first and second laws of thermodynamics (Beard *et al.*, 2002, 2004), or metabolite concentrations (Kümmel *et al.*, 2006b; Henry *et al.*, 2007). Nevertheless, these extensions require the inclusion of additional experimental data and may result in more complex mathematical formulation hindering their practical use.

Some predictions of constraint-based models may be wrong in cases where modeling assumptions do not hold. For instance, some metabolites do accumulate in the cell, and the mass balance assumption clearly does not hold for these. In general, the concentration of specific metabolites may be high enough relatively to the fluxes they are involved in for the mass balance approximation to become clearly false.

In practice, many of the analytical methods that have been developed for constraint-based models focus on defining and characterizing sets of feasible flux distributions. Others focus on a single distribution. The diversity of flux distributions compatible with constraints in a given environment can be viewed as reflecting the diversity of the metabolic states the cell may find itself in. Nevertheless, the space of feasible flux distributions features biologically informative properties whose determination requires adequate techniques; these will be introduced in the next sections of this review.

Building the models

The level of detail necessary to build a constraint-based model of a bacterium's metabolism is relatively low; the only information required is the precise reaction stoichiometries and directions, in order to account for mass balance and irreversibility constraints. To reflect the global biochemical capabilities of the organism, the model also needs to encompass the complete set of metabolic activities that can occur within it – or a reasonable approximation thereof. This comprehensiveness requirement and the high number of metabolic reactions make the actual construction of such models a challenging task in itself. In this section, we will review the main methods and resources helping in this task. We will first show how information from genome annotation can be used to infer biochemical reactions at large scale, a task commonly called *metabolic network reconstruction*. We will then review the techniques commonly used to assess the consistency of reconstructed models, and show how missing biochemical activities can be identified to complete the model.

Initial reconstruction of metabolic models

The most reliable evidence from which the presence of a metabolic reaction in a species can be inferred is experimental proof of the respective biochemical activity. Such biochemical results have been accumulated for several decades, mostly from dedicated experiments targeting well-defined activities. As a consequence, the corresponding reactions have often been precisely and reliably characterized. Exploiting these results to reconstruct the whole metabolism of an organism is a labor-intensive task, however, as it requires processing a high volume of literature. Most existing metabolic models have been reconstructed in this manner and for extensively studied organisms. For instance, the most complete bacterial model available to date – namely iAF1260, the latest model of *Escherichia coli* metabolism – includes references to more than 320 articles (Feist *et al.*, 2007). Two types of databases centralize biochemical knowledge: enzyme-centric ones, which collect functional information acquired on enzymes, for example BRENDA (Barthelmes *et al.*, 2007) or SwissProt (Boutet *et al.*, 2007); and pathway databases, aimed at describing the biochemistry of metabolic processes, for example EcoCyc for *E. coli* metabolism (Karp *et al.*, 2007) or UM-BDD for microbial biodegradation pathways (Ellis *et al.*, 2006) (see Table 1).

These biochemical clues are typically incomplete relatively to the set of all possible activities, especially for less studied organisms. In addition, while technologies aiming at high-throughput characterization of biochemical activities are improving, they are not yet mature enough to provide reasonably good coverage. Genes corresponding to enzymes that have been experimentally characterized have nevertheless been identified. Their homologues in the genome of such species can be identified using comparative genomics methods, thereby indicating the presence of the associated biochemical activities.

The traditional path to inferring metabolic reactions from the genome of an organism is gene-centric, at least in its first steps. Nearly all available genome sequences are now systematically processed through automated annotation pipelines, which identify coding sequences and infer functional annotations. Covering all relevant methods would be beyond the scope of this article, but thorough reviews can be found elsewhere (Médigue & Moszer, 2007). Basically, coding sequences are first identified using highly efficient gene-finding algorithms [such as GENEMARK (Besemer *et al.*, 2001), GLIMMER (Delcher *et al.*, 1999), or AMIGENE (Bocs *et al.*, 2003)], which discard the ORFs that are not likely to be coding for a protein. Functional annotations are then sought for each gene using complementary approaches: sequence homology with proteins of known function [stored for instance in UniprotKB (UniProt, 2008)], conservation of genomic structure with annotated species (e.g. synteny), and prediction of functional domains (Apweiler *et al.*, 2000; Claudel-Renard *et al.*, 2003).

Table 1. Data sources for metabolic model reconstruction and refinement

DNA sequence and genome annotation databases		
DDBJ	http://www.ddbj.nig.ac.jp/	General nucleotide sequence database
EMBL	http://www.ebi.ac.uk/embl/	General nucleotide sequence database
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/	General nucleotide sequence database
Integr8	http://www.ebi.ac.uk/integr8/	Integrated information on complete genomes
CMR	http://cmr.jcvi.org/	Integrated information on complete prokaryotic genomes
IMG	http://img.jgi.doe.gov/	Integrated system for analysis and annotation of microbial genomes
SEED	http://seed-viewer.theseed.org/	Integrated system for analysis and annotation of genomes using functional subsystems
Protein and enzyme databases		
BRENDA	http://www.brenda-enzymes.info/	Comprehensive enzyme information system gathering data collected from the literature by curators
ENZYME	http://www.expasy.ch/enzyme/	Enzyme nomenclature database providing extensive information on all enzymes with an associated EC number
UniProt	http://www.ebi.ac.uk/uniprot/	Universal Protein Resource gathering protein sequences and annotations from SwissProt (manually reviewed), trEMBL (computer annotated), and PIR
TransportDB	http://www.membranetransport.org/	Predictions of membrane transport proteins for fully sequenced genomes
PSORTdb	http://db.psort.org/	Repository of experimentally determined and predicted protein localizations
Prolinks	http://prolinks.mbi.ucla.edu/	Database of predicted functional links between proteins
STRING	http://string.embl.de/	Database of known and predicted protein–protein interactions
Metabolic databases		
CheBI	http://www.ebi.ac.uk/chebi/	Database on small molecules of biological interest
Pubchem	http://pubchem.ncbi.nlm.nih.gov/	Database on small molecules
LipidMaps	http://www.lipidmaps.org/	Database on lipid metabolites
Reactome	http://www.reactome.org/	Curated database of biological pathways
KEGG	http://www.genome.jp/kegg/	Suite of databases comprising information on compounds, reactions, pathways, genes/proteins
BioCyc	http://www.biocyc.org/	Collection of organism-specific pathway/genome databases, including a curated multiorganism pathway database: MetaCyc
UniPathway	http://www.grenoble.prabi.fr/obiwarehouse/unipathway/	Curated resource of metabolic pathways linked to UniProt enzyme database
UM-BBD	http://umbbd.msi.umn.edu/	Database on microbial biocatalytic reactions and biodegradation pathways
Experimental data repositories		
IntAct	http://www.ebi.ac.uk/intact/	Repository of reported protein interactions
DIP	http://dip.doe-mbi.ucla.edu/	Database of experimentally determined interactions between proteins
Array Express	http://www.ebi.ac.uk/aerep/	Public repository of microarray data
GEO	http://www.ncbi.nlm.nih.gov/geo/	Public repository of microarray data
ASAP	http://asap.ahabs.wisc.edu/	Repository of results of functional genomics experiments for selected bacterial species
E. coli multi-omics DB	http://ecoli.iab.keio.ac.jp/	Comprehensive dataset of transcriptomic, proteomic, metabolomic, and fluxomic experiments for *E. coli* K12
Systomonas	http://www.systomonas.de/	Repository of 'omics' datasets and molecular networks for pseudomonads species
PubMed	http://www.pubmed.org/	Database on biomedical literature
Metabolic model repositories		
BiGG	http://bigg.ucsd.edu/	Repository of reconstructed genome-scale metabolic models
BioModels	http://www.ebi.ac.uk/biomodels/	Database of mathematical models of biological systems

Combining the above methods and information sources increases the reliability of the annotation transfers from proteins of known function to new genes. Current annotation pipelines succeed at assigning a function to 50–80% of the genes (Serres *et al.*, 2004). A number of databases provide such automatically generated annotations for most sequenced bacterial genomes (see Table 1).

In order to build a metabolic model, it is necessary to identify the specific chemical conversions catalyzed by each enzyme, together with the corresponding stoichiometries. Functional annotations of enzymes therefore need to be translated into appropriate chemical equations. The Enzyme Commission (EC) numbers classification offers an unambiguous way to identify enzyme-catalyzed reactions. When provided by the enzyme annotations, these numbers directly specify which reactions they catalyze. Several enzyme and metabolic databases provide the correspondence between EC numbers and reactions (see Tables 1 and 2). These

Table 2. Type of information provided by each data source

Type of information	DDBJ	EMBL	GenBank	Integr8	CMR	IMG	SEED	BRENDA	ENZYME	UniProt	TransportDB	PSORTdb	ProLinks	STRING	ChEBI	PubChem	LipidMaps	Reactome	KEGG	BioCyc	UniPathway	UM-BBD	IntAct	DIP	Array Express	GEO	ASAP	E. coli multi-omics DB	Systomonas	PubMed
Biochemical activities	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Enzyme specificity	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Enzyme localization	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Reaction equation	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Reaction direction	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Metabolite formula	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
GPR association[1]	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Biomass composition	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Experimental observations	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

metabolic databases are often comprehensive catalogues of known biochemical reactions with the associated chemical information, including stoichiometry: they include most of the reference knowledge needed to build metabolic models.

Several issues hinder this translation process. First, enzymatic activities that have been identified only recently are usually not included in the EC classification. Furthermore, full EC numbers are not always systematically assigned in the annotation process. As a result, many annotations retrieved from protein databases are only textual (as in UniProtKB) or ontology based [as in Gene Ontology (Ashburner*et al.*, 2000)] and do not provide the required metabolic information directly. To address this shortcoming, PATHOLOGIC, the metabolic network reconstruction software tied to the BioCyc metabolic databases, includes an algorithm performing the identification of gene-reaction links from textual annotations (Karp *et al.*, 2002) (see Table 3). This procedure relies on a dictionary of synonyms, however, and may fail at recognizing an explicit reaction when uncommon terms are used. An expert curation step is thus necessary, for which metabolic pathway databases provide useful guidance. Recent initiatives specifically aim at solving this issue: for instance, textual annotations in UniProtKB/SwissProt are being progressively replaced by direct references to reactions from UniPathway, a metabolic database in which all reaction steps are specified up to the chemical level (see Table 1).

The broad specificity of some enzymes may also significantly increase the number of distinct reactions they can catalyze. For instance, enzymes annotated with alcohol dehydrogenase activity (EC 1.1.1.1) may catalyze the degradation of several distinct alcohols. Similarly, enzymes acting on lipids are often not specific to the length of their carbon chain. In such cases, functional annotations often report the activity using generic metabolites (e.g. 'an alcohol' or 'a fatty acid') representing the entire set of possible substrates. Instantiating reactions with specific metabolites is required when building a metabolic model, however, as accounting for the mass balance constraint requires that all metabolites should be well defined. It is thus necessary to identify for each generic compound the corresponding set of specific compounds, as much for primary substrates as for cofactors. This task is complicated by the combinatorial effect, because the number of substrate combinations may significantly increase the number of specific reactions. To address this issue, databases of chemical species can be used to identify all metabolites of a given chemical category (see Tables 1 and 2). In order to determine which metabolites are preferentially recognized by enzymes, processing the literature or browsing information collected in enzyme databases such as BRENDA (Barthelmes *et al.*, 2007) is often necessary. Metabolites involved in metabolic pathways that have already been inferred may also help in selecting the most relevant substrates.

Table 3. Methods for model reconstruction

Metabolic model reconstruction (beyond the use of dedicated metabolic databases)	
Identification of metabolic reactions from textual gene annotations	Karp *et al*. (2002)
Direct inference of metabolic reactions from genome sequence	Sun & Zeng (2004), Arakawa *et al*. (2006), Notebaart *et al*. (2006)
Use of metabolic context to complete pathways	Karp *et al*. (2002), Arakawa *et al*. (2006), DeJongh *et al*. (2007)
Metabolic model consistency checks	
Flux variability analysis: identification of reactions that are predicted to never carry any flux	Mahadevan & Schilling (2003)
Identification of dead-end metabolites, which can never be produced or consumed.	Segrè *et al*. (2003), Ebenhöh *et al*. (2004), Imielinski *et al*. (2005), Kumar *et al*. (2007)
Assessment of thermodynamic consistency and assignment of reaction directions.	Yang *et al*. (2005), Kümmel *et al*. (2006a, b)
Gap filling and model expansion	
Graph-based metabolic network expansion using shortest metabolic paths	Arita (2003), Boyer & Viari (2003)
GapFill: optimization-based network expansion and reaction reversibility changes to solve dead-end metabolite inconsistencies	Kumar *et al*. (2007)
Optimization-based metabolic network expansion to resolve inconsistent growth phenotypes	Reed *et al*. (2006a, b)
Network-based identification of candidate genes for orphan metabolic activities	Osterman & Overbeek (2003), Green & Karp (2004), Chen & Vitkup (2006), Kharchenko *et al*. (2006), Fuhrer *et al*. (2007)

Alternative approaches to metabolic network reconstruction bypass the classical annotation step altogether, taking instead advantage of the curated links between enzyme-encoding gene sequences and reactions [or EC numbers, as in the Genome-Based Modeling (GEM) system (Arakawa *et al*., 2006)] provided by some metabolic databases. Orthology relationships are sought between reference sequences from these databases and the coding sequences from the new genome. While these methods [e.g. AUTOGRAPH (Notebaart *et al*., 2006), or IDENTICS (Sun & Zeng, 2004), see Table 3] simplify the reconstruction process, they usually do not benefit from advanced annotation techniques, such as those derived from structural genomics or domains recognition, and are more difficult to combine with expert annotation. They are also conditioned on the availability of curated gene-reaction associations for a set of reference organisms.

The reconstruction of the metabolism of a new organism can also benefit from the knowledge of complete pathways in related organisms. Metabolic databases often group reactions into pathways or modules that indicate known co-occurrence relationships between reactions that hold across several organisms. Three main resources provide this type of information: MetaCyc (Caspi *et al*., 2006), KEGG Modules (Kanehisa *et al*., 2007), and SEED (Overbeek *et al*., 2005) (see Tables 1 and 2). Metabolic model reconstruction procedures tied to such databases can exploit the known co-occurrences of reactions across reference organisms whose metabolism has been extensively studied (Arakawa *et al*., 2006). An instance of a reconstruction procedure taking advantage of this notion of metabolic context is again PATHOLOGIC, which infers the presence of pathways rather than that of single reactions when possible. A reconstruction procedure based on the SEED database was also proposed recently (DeJongh *et al*., 2007); it includes a check that the inferred pathways can be properly connected to form a 'working' model. By leveraging a specific form of 'guilt-by-association,' approaches of this type may be able to retrieve reactions catalyzed by enzymes that cannot be correctly identified using current methods. In addition, the presence of spontaneous reactions in the organism may be identified by the occurrence of neighboring reactions in reference metabolic pathways.

In addition to their equations, the reversibility and localization of reactions need to be determined for metabolic models. Few metabolic or enzyme databases report on the reversibility of reactions in *in vivo* conditions (see Table 2). When not found in the literature, reversibility is therefore often determined using simple thermodynamic considerations based on the reaction Gibbs energy, if it is known, or on basic rules depending on the energy equivalents (e.g. NADH or ATP) involved in the reactions (Ma & Zeng, 2003; Kümmel *et al*., 2006a). Even though very few compartments divide bacterial cells (with periplasm and cytoplasm as the only main compartments in gram-negative bacteria), the presence of such physical separation between metabolites need to be included in their metabolic models. Enzymes present in one compartment cannot interact with metabolites present in another one. To properly model the effect of compartments, the localization of enzymes and the transport of metabolites need to be determined. Information on the localization of enzymes and reactions is seldom included in metabolic databases. Curated versions of BioCyc databases, especially MetaCyc, are a welcome exception, however (Caspi *et al*., 2006). When not found in the literature, localization can be inferred using *ab initio* predictions from enzyme sequences (Schneider & Fechner, 2004),

or determined experimentally, for example using fluorescence microscopy (Meyer & Dworkin, 2007). Transport of metabolites can be inferred using comparative genomics tools that identify transport enzymes [e.g. TransportDB (Ren *et al.*, 2004)]. Yet, such methods hardly determine the specificity of transporters; knowledge of transported metabolites is therefore often completed using direct information on the microorganism's physiology and the metabolites it was shown to utilize in growth experiments.

Overall, reconstructing a constraint-based model for an organism's metabolism involves collecting various types of information. A summary of the respective contributions of each data source to the model construction is shown in Table 2.

Checking the consistency of reconstructed models

Once a draft metabolic model is obtained, its consistency can be checked using a set of simple tests (see Fig. 2): is the model chemically and physically coherent? Are there remaining 'dead-ends' in metabolic pathways or reactions bound to be inactive? Is the model able to produce essential metabolites from a known growth medium?

Constraint-based metabolic models fundamentally rely on reaction stoichiometries to properly account for the mass balance in metabolism at steady state. It is therefore crucial that all chemical equations are correctly balanced to avoid unrealistic creation or destruction of matter. To control the correctness of the reaction stoichiometries, the atom balance of each reaction can be checked using the chemical formulae of the metabolites, which are typically found in databases of chemical compounds (see Table 1). For cases where the formula is not available for all metabolites, a method was recently introduced to detect such balance errors in metabolic models by solely comparing chemical equations – for instance, reactions A → B and A → B+C would be identified by this method as 'stoichiometrically inconsistent,' because balancing both equations would require that at least one of the metabolites has a null or negative mass (Gevorgyan *et al.*, 2008).

The assumptions on which constraint-based models are founded do not enforce thermodynamic consistency on the fluxes. Flux distributions obeying conservation of mass can still include internal cycles that violate thermodynamic laws, allowing for instance the artificial generation of high-energy cofactors. To prevent models from predicting such unrealistic metabolic modes, extensions of the modeling framework were proposed that directly enforce these laws (Beard *et al.*, 2002). Their nonlinear nature entails costly computations, however, which hinder the use of such modeling extensions in practice. In order to provide thermodynamically consistent models without including such extensions, methods have been developed to detect inconsistent cyclic modes in draft metabolic models, and propose changes in reaction reversibility that would avoid those modes from being predicted (Yang *et al.*, 2005; Kümmel *et al.*, 2006a).

Before one can reap the benefits of having a model, the model should be functional, i.e. it should be checked that non-null fluxes can actually be predicted. This relates to the completeness of the model, because for instance a missing

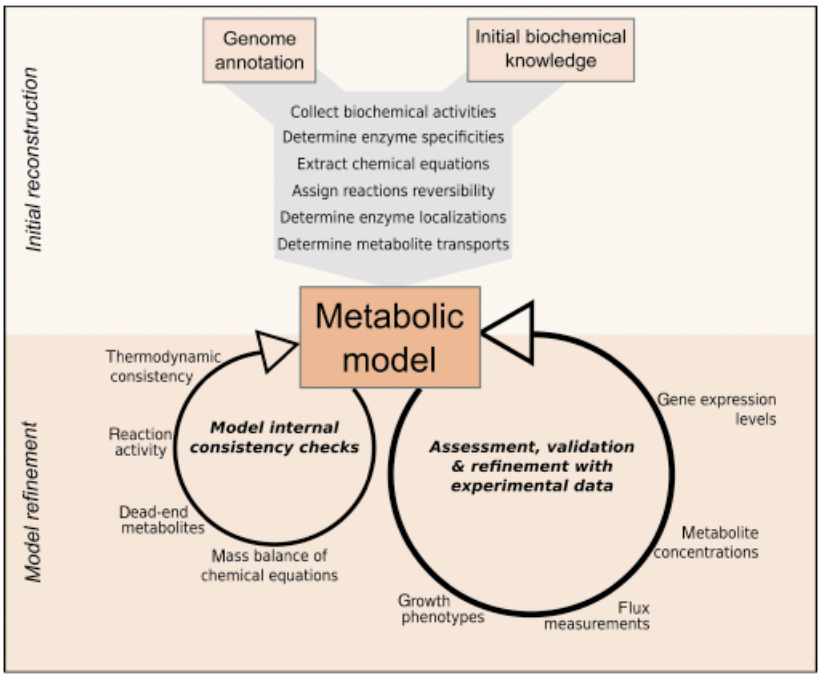


Fig. 2. Pipeline for model reconstruction and refinement. An initial model is reconstructed from genome annotations and from preexisting knowledge on the species' biochemistry and physiology. Besides collecting the biochemical activities, this task includes several additional key steps. The resulting model is then iteratively corrected and refined, according to internal consistency criteria and by comparing its predictions to experimental data.

reaction in a linear pathway would prevent any non-null flux from being predicted in it at steady state, thereby inactivating all other reactions in the pathway. Metabolites that are never consumed or never produced, so-called 'dead-ends,' typically reveal that reactions are missing. In order to help investigate and correct these so-called 'metabolic gaps,' methods have been developed that assess whether reactions can be active in the model (Reed & Palsson, 2004), identify dead-end metabolites (Kumar *et al.*, 2007) or directly track the producibility of metabolites from source metabolites (Segrè *et al.*, 2003; Ebenhöh *et al.*, 2004). In case the model is later used to predict growth phenotypes (see Applications of metabolic models), the producibility of biomass precursors and the completeness of their biosynthetic pathways should be especially checked beforehand. Dedicated procedures have been designed to systematically perform these checks for newly reconstructed models (Segrè *et al.*, 2003; Imielinski *et al.*, 2005; Senger & Papoutsakis, 2008). Solving such inconsistencies often involves filling metabolic gaps or completing the network with additional metabolic pathways.

The methods presented in this section check the consistency of the reconstructed model with respect to a set of basic rules (see Table 3). We will review in the section on model applications how model predictions can also be confronted with experimental data, providing consistency checks of the model with respect to diverse additional experimental evidence. Interpreting and solving identified inconsistencies of either type are key to improving the quality of the metabolic model.

Targeted searches for missing metabolic activities

Consistency checks (either internal to the model or relative to experimental datasets) may showthat the reconstructed model is incomplete and lacks some metabolic reactions. Resolving these metabolic gaps entails expanding the model by identifying and including missing biochemical activities. This process basically consists of two steps: (1) identifying plausible candidate reactions that could complete the model and (2) finding genes that could catalyze the hypothesized activities.

Reactions contained in metabolic databases are the primary source of information for completing the metabolic model (see Table 1). The search for candidate reactions within these databases can be facilitated using knowledge of existing pathways (as in MetaCyc, SEED, or UM-BBD, see Table 1) or computational methods (Arita, 2003; Boyer & Viari, 2003; Kumar *et al.*, 2007) (see Table 3). In the latter category, the GapFill method was specifically developed to identify dead-ends in models, and correct them by adding reaction from a global repository of reactions, changing the reversibility status of reactions, or adding transporters (Kumar *et al.*, 2007). The addition of reactions to the model is guided by an optimization step minimizing the number of reactions. Similarly, Reed *et al.* (2006b) proposed a method which drives the expansion of the metabolic model to account for the utilization of additional external compounds. For metabolites experimentally shown to be used by the organism but not predicted as such by the model (see Applications of metabolic models on growth phenotype predictions for methods to perform these predictions), their method automatically proposes minimal sets of reactions from a repository of reactions that, if added, would allow the model to exploit the external metabolites.

The set of reactions referenced in metabolic databases is far from being comprehensive: the right candidates for completing the model may not yet be known. Computational and experimental approaches have been proposed to extend this 'universe of possible reactions.' On the computational side, several methods originating from the field of chemo-informatics have been designed to infer chemical transformations (Gasteiger, 2005). Some of them have been more specifically adapted to biochemical transformations, using rules on enzymatic conversions to infer new conversions for biologically relevant metabolites (Klopman *et al.*, 1994; Arita, 2000; Hatzimanikatis *et al.*, 2005; Ellis *et al.*, 2008).

Numerous experimental methods are also being developed to explore the range of possible biochemical reactions. MS and nuclear magnetic resonance (NMR) techniques are able to identify and quantify large sets of metabolites at high throughput (Dunn *et al.*, 2005; Dettmer *et al.*, 2007). Computational methods have been proposed to infer reactions from MS data, by analyzing mass differences between related metabolites (Breitling *et al.*, 2006) or correlations between metabolite concentrations across distinct conditions (Steuer, 2006). They do not provide direct evidence for biochemical transformations, however: their predictions should be treated as clues to be confirmed by additional information. Although mostly used to determine metabolic fluxes, atom-labeling experiments could also become powerful tools to elucidate novel metabolic pathways (Sauer, 2006). They can advantageously complement computational *ab initio* pathway inference methods by selecting candidate pathways that are compatible with observed isotopic patterns. Finally, untargeted enzyme activity screenings have recently been performed to identify the substrates of enzymes of unknown function and discover novel activities (Saghatelian *et al.*, 2004; Saito *et al.*, 2006). The availability of large-scale libraries of ORF clones (Kitagawa *et al.*, 2005) should increase the likelihood of such methods expanding the store of known reactions.

The search for candidate genes for orphan metabolic activities is in some ways the reverse of the classical genome annotation problem (i.e. searching the function of identified genes). Yet, many of the tools developed to determine gene functions can be adapted for this purpose. Sequence homology to already characterized genes is central to most methods for candidate gene detection, but combining it

with additional types of evidence can significantly improve performance. For instance, several approaches exploit functional links, such as gene neighborhood, gene co-expression, protein interaction, or phylogenetic co-occurrence, to relate candidate genes with genes involved in the same metabolic pathways or close in the metabolic network (Osterman & Overbeek, 2003; Green & Karp, 2004; Chen & Vitkup, 2006; Kharchenko *et al.*, 2006; Fuhrer *et al.*, 2007). Databases such as STRING (von Mering *et al.*, 2007) or Prolinks (Bowers *et al.*, 2004) compile large sets of functional links across a wide range of organisms. On the experimental side, enzyme activity screenings are used to validate the generated candidates. Furthermore, when the orphan activity is associated to a specific phenotype, screens of systematic knockout mutant phenotypes can help in identifying candidates (Aghaie *et al.*, 2008).

The two types of methods – finding candidate reactions or candidate genes – benefit from being used in combination, as identifying genes for putative reactions can help in selecting the proper reactions to include.

Applications of metabolic models

A wealth of computational methods has been developed to help analyze biological properties revealed by reconstructed metabolic models. Not only would a comprehensive and technical description exceed the scope of this review, but these methods have been extensively covered elsewhere, either on the technical side (Price *et al.*, 2004) or for applications on a specific organism, i.e. *E. coli* (Feist & Palsson, 2008). We will provide here the reader with a review on the main applications for which constraint-based models have been most successful and are mostly promising for bacterial species. We will distinguish four main types of applications: (1) analysis of network properties of metabolism, (2) prediction and analysis of bacterial growth phenotypes, (3) model-based interpretation of experimental data, and (4) metabolic engineering.

Analysis of network properties

The principle of constraint-based modeling consists in studying the set of reaction fluxes – namely flux distributions – that are achievable at steady state given the constraints imposed on the system. Reaction fluxes can vary inside a continuous set of possible values. This set can encompass significant variability at the level of individual pathway or reaction fluxes. A wide range of methods have been designed to explore that variability and analyze specific properties of metabolites and reactions which emerge from the flux constraints.

One approach consists in sampling the set of achievable flux distributions (Almaas *et al.*, 2004; Reed & Palsson, 2004; Wiback *et al.*, 2004). Methods that provide a uniform sampling of the possible states have been proposed (Almaas *et al.*, 2004; Wiback *et al.*, 2004). By sampling a significant number of metabolic states, these approaches offer an overview of the range of flux distributions that can occur in the metabolic network at steady state. The 'uniform' nature of the sampling is based only on the mathematical description of the set of possible flux distributions, avoiding any prior assumption on which metabolic states are most likely to be selected *in vivo*. For instance, these sampling methods have been used to evaluate the relative occurrence of reactions within the set of possible flux distributions and across several environmental conditions (Almaas *et al.*, 2004). This analysis showed that a few reactions are active in many sampled flux distributions and carry high fluxes – forming a so-called high-flux metabolic backbone – while many others are active in few sampled flux distributions and carry low fluxes. Similar methods were also used to evaluate the correlation of flux values between pairs of reactions across sampled metabolic states (Reed & Palsson, 2004; Becker *et al.*, 2007) and thereby determine metabolic dependencies between reactions. From a more theoretical angle, sampling was also used to evaluate the size of the set of possible flux distributions (Wiback *et al.*, 2004; Braunstein *et al.*, 2008). When computed for distinct (genetic perturbation× environmental condition) pairs, the relative sizes of the corresponding flux distribution sets were interpreted as indicators of the respective diversity of metabolic states in the tested conditions (Wiback *et al.*, 2004).

The diversity of achievable metabolic fluxes can also be evaluated locally for each reaction. Flux variability analysis was designed for this purpose: an optimization procedure computes the minimal and maximal allowed flux of each reaction independently (Mahadevan & Schilling, 2003). This procedure identifies reactions that do not carry any flux, or conversely those that carry non-null flux in all possible metabolic states. Flux variability analysis has been broadly used to predict the activity of reactions for specific sets of metabolic constraints (Mahadevan & Schilling, 2003; Reed & Palsson, 2004; Teusink *et al.*, 2006; Feist *et al.*, 2007; Henry *et al.*, 2007; Shlomi *et al.*, 2007a).

Flux sampling or flux variability approaches only provide partial description of the set of possible flux distributions. To get a comprehensive picture of the possibilities, methods which compute elementary modes (Schuster *et al.*, 2000) and extreme pathways (Schilling *et al.*, 2000) have been developed. These notions differ only slightly in their mathematical formulation (Klamt & Stelling, 2003; Papin *et al.*, 2004): the main idea is to determine the set of elementary and independent metabolic routes that can occur in the metabolic model. These elementary routes are flux distributions that (1) respect all assumed constraints, including steady state and irreversibility, and (2) are elementary in the sense that they are composed of a minimal set of active reactions. This second condition ensures that the flux

distribution is not decomposable into a combination of smaller elementary routes. It can be shown that any achievable flux distribution can be expressed as a combination of such elementary routes. This property, together with the fact that the set of elementary routes is unique, independently of the method used to compute it (Klamt & Stelling, 2003), has inspired numerous applications. This subfield is also known as metabolic pathway analysis. For instance, elementary modes and extreme pathways have been used to exhaustively describe the independent metabolic routes occurring in newly reconstructed models, often sorted by metabolic function (Schilling & Palsson, 2000; Van Dien & Lidstrom, 2002; Papin *et al.*, 2002). The redundancy of routes can be assessed and the respective yields of routes of conversion can be compared (Papin *et al.*, 2002). Conversely, the relative importance of reactions in metabolism was scored using elementary routes, reactions involved in many routes being likely to be key players in metabolism (Stelling *et al.*, 2002). Finally, metabolic dependencies between reactions which are stronger than those determined only by analyzing the correlation of fluxes in sampled distributions can be deduced from knowing elementary routes. Reactions that always appear jointly in elementary routes are bound to operate together (Pfeiffer *et al.*, 1999). The main obstacle in metabolic pathway analysis is the size and complexity of the metabolic models, as the number of elementary routes dramatically increases with the size of the model (Yeung*et al.*, 2007). The computation of all routes is currently only tractable for medium-size models, although significant progresses have been made recently (Terzer & Stelling, 2008).

Alternative approaches have been developed in order to explore metabolic dependencies in models of larger size. One of them, flux coupling analysis, has become a popular analytical tool (Burgard *et al.*, 2004). Flux coupling analysis identifies all pairs of reactions whose fluxes are always coupled at steady state. It has been used in a wide range of studies, and the resulting sets of coupled reactions were for instance compared with correlations observed in the transcriptional states of enzymes (Reed & Palsson, 2004; Notebaart *et al.*, 2008), interpreted with respect to the structure of the metabolic regulation (Notebaart *et al.*, 2008), and used to study the horizontal transfer of genes during bacterial evolution (Pál *et al.*, 2005a, b). Similar methods were developed to study metabolic relationships between metabolites, either by simply examining the co-occurrence of metabolites in reactions (Becker *et al.*, 2006) or by determining conservation relations between metabolites (Nikolaev *et al.*, 2005; Imielinski *et al.*, 2006). This last type of method was applied to determine coupling relationships between metabolite concentrations, identify metabolite pools sharing conserved chemical moieties (Nikolaev *et al.*, 2005), and exhaustively predict distinct minimal growth media for *E. coli* (Imielinski *et al.*, 2006).

Prediction of growth phenotypes

One of the primary uses of genome-scale metabolic models is the prediction of growth phenotypes (Price *et al.*, 2004; Palsson, 2006). Because these models aim at comprehensiveness, they are able to account for all main metabolic processes contributing to growth, i.e. the production of energy and biomass precursors from external metabolites. Growth phenotypes can therefore be predicted by examining to which extent metabolic requirements for growth, in terms of energy generation and biomass precursors synthesis, can be fulfilled by flux distributions from the model. Growth phenotypes can be predicted either in a qualitative manner (prediction of the mere ability to grow) by checking piecemeal for the producibility of each biomass precursor metabolite (Imielinski *et al.*, 2005), or in a quantitative manner (prediction of growth performance) by including a biomass reaction consuming them in proportion to their ratio in biomass composition and studying the flux values it can attain (Price *et al.*, 2004). Determining biomass composition is therefore a necessary prerequisite to growth phenotype predictions. This is often achieved by examining the relevant literature or adapting known biomass compositions of related organisms. The Flux Balance Analysis (FBA) method was specifically designed to predict quantitative growth phenotypes (Varma & Palsson, 1994b; Price *et al.*, 2004). It computes the maximal growth yield achievable in the metabolic model by maximizing the biomass reaction flux (representing the growth rate) given a set of bounded intake rates for external substrates. FBA relies on the strong assumption that bacteria have optimized their growth performance in a subset of possible environments during their evolution, thereby making the maximization of biomass production a driving principle for metabolic operation (Varma & Palsson, 1994b). This assumption has been confirmed by experiments in several cases (Edwards *et al.*, 2001). Using FBA, global quantitative relationships can be predicted between the input rates of nutrients, the output rates of byproducts, and the growth rate (Stephanopoulos *et al.*, 1998; Edwards *et al.*, 2002; Price *et al.*, 2004).

The global energy consumption of the cell can significantly influence the outcome of quantitative growth phenotype predictions. Two ATP hydrolysis fluxes are added to the models in order to properly account for it. One is constant and models the non-growth-associated maintenance, which represents the fraction of the energy demand necessary for the cell survival that is independent from its growth rate, for example to maintain the right ionic strength (Stouthamer & Bettenhaussen, 1973). The second flux is proportional to the growth rate and corresponds to the energy demand associated with growth beyond the mere requirements of metabolic pathways – which are already directly accounted for in the model – for example energy for cell division or assembly

Table 4. Existing genome-scale metabolic models for bacterial organisms

					Experimental assessment		
Organism	Reference	Genes	Reactions*	Metabolites†	Wild-type growth phenotypes	Knockout mutant growth phenotypes	Quantitative growth measures
Acinetobacter baylyi	Durot *et al*. (2008)	774	875	701	173/190 (91%)	1138/1208 (94%)	–
Bacillus subtilis	Oh *et al*. (2007)	844	1020	988	200/271 (74%)	720/766 (94%)	–
Clostridium acetobutylicum	Lee *et al*. (2008a)	432	502	479	10/11 (91%)	–	X
Clostridium acetobutylicum	Senger & Papoutsakis (2008)	474	552	422	–	–	–
Escherichia coli‡	Feist *et al*. (2007)	1260	2077	1039	129/170 (74%)	1152/1260 (92%)	X
Geobacter sulfurreducens	Mahadevan *et al*. (2006)	588	523	541	–	–	X
Haemophilus influenza	Schilling & Palsson (2000)	412	461	367	–	–	–
Helicobacter pylori§	Thiele *et al*. (2005)	341	476	485	–	54/72 (75%)	–
Lactobacillus plantarum	Teusink *et al*. (2006)	721	643	531	–	–	X
Lactococcus lactis	Oliveira *et al*. (2005)	358	621	422	–	–	X
Mannheimia succiniciproducens	Hong *et al*. (2004)	335	373	332	–	–	–
Mycobacterium tuberculosis	Beste *et al*. (2007)	726	849	739	–	547/705 (78%)	X
Mycobacterium tuberculosis	Jamshidi & Palsson (2007)	661	939	828	–	132/237 (56%)	X
Neisseria meningitidis	Baart *et al*. (2007)	555	496	471	–	–	X
Pseudomonas aeruginosa	Oberhardt *et al*. (2008)	1056	883	760	78/95 (82%)	893/1056 (85%)	–
Pseudomonas putida	Nogales *et al*. (2008)	746	950	710	84/90 (93%)	665/746 (89%)¶	X
Rhizobium etli	Resendis-Antonio *et al*. (2007)	363	387	371	–	–	–
Staphylococcus aureus	Becker & Palsson (2005)	619	641	571	–	–	–
Staphylococcus aureus	Heinemann *et al*. (2005)	551	774	712	–	8/14 (57%)	–
Streptomyces coelicolor	Borodina *et al*. (2005)	700	700	500	54/58 (93%)	11/12 (92%)	X

First two columns of experimental assessment show the number of correct predictions among all experimentally determined qualitative growth phenotypes. Last column specifies whether the model has been assessed against quantitative growth rate measurements.

*Number of distinct reactions including transport processes.

†Number of biochemically distinct metabolites.

‡This model is an update of two earlier models for *E. coli* (Edwards & Palsson, 2000; Reed *et al*., 2003).

§This model is an update of an earlier model for *H. pylori* (Schilling *et al*., 2002).

¶Using gene essentiality data for *Pseudomonas aeruginosa*.

of higher order cell structures. These two parameters are usually determined by fitting growth yield predictions derived using FBA to measured growth yields provided by growth monitoring experiments (Reed *et al*., 2006a). Measurements of growth yields for distinct growth rates are sufficient to fit both growth-associated and non-growth-associated maintenance parameters (Varma & Palsson, 1994a). The values of these parameters were determined using experimental growth measurements for a significant proportion of reconstructed models (see Table 4).

Once fitted, and assuming these parameters remain constant across environments, the model can be used to predict growth rates on different media (Edwards *et al*., 2001). Predicted growth yields revealed to be consistent with observed ones on a significant number of media for *E. coli* (Edwards *et al*., 2001). Inconsistencies between predicted and observed growth yields can have multiple interpretations. First, the assumption of optimal substrate utilization can be questionable for growth predictions on environments that are not commonly encountered by the organism (Ibarra *et al*., 2002; Schuster *et al*., 2008). Using an adaptive evolution experiment on *E. coli* cells grown in glycerol minimal medium, Ibarra and colleagues actually observed that, while the initial growth yield was suboptimal, it progressively evolved to reach the optimal value predicted by the model. Other biological constraints, such as regulation or capacity constraints, may also prevent the organism from using optimal flux distributions (Oliveira *et al*., 2005; Feist *et al*., 2007). Comparing predictions of growth phenotypes with experimental measures may also help in refining the model. A model component that is often refined using quantitative growth predictions is the stoichiometry of proton translocation that occurs in reactions of electron transport systems, such as the respiratory chain. These stoichiometries are often hard to determine *a priori*, yet they impact directly the P/O ratio and the efficiency of energy generation (Reed *et al*., 2006a). With the help of a metabolic model and growth yield measurements on several distinct media, Feist *et al*. (2006) studied the unknown proton translocation stoichiometry of such a reaction in *Methanosarcina barkeri* by determining for each media the model maintenance parameters that provided the best growth yield

predictions for different hypothesized values of the stoichiometry. Assuming that maintenance should not significantly change across media, they selected the stoichiometry that triggered the smallest variation among the determined maintenance parameters across the environments.Other studies investigated the stoichiometry of proton translocation in the respiratory chain by directly exploiting measured ratios of electron acceptor (e.g. oxygen, or Fe(III) in *Geobacter sulfurreducens*) consumption rate vs. carbon source consumption rate and growth rate (Heinemann *et al.*, 2005; Mahadevan *et al.*, 2006).

Models can readily predict the effect of gene deletion on growth phenotypes. To that end, a layer of Gene Protein Reaction associations – usually called GPR (Reed *et al.*, 2003) – is added to the model to predict the effect of gene deletion on reaction activity. Each reaction is associated to its enzyme-encoding genes by a Boolean rule: genes encoding for subunits of an enzymatic complex are linked with an AND rule, while genes encoding for alternative enzymes are linked with an OR rule. Using GPR rules, gene deletions are translated into 'blocked' reactions, which are then inactivated in the model by constraining their fluxes to zero. FBA can be applied to predict growth phenotypes of gene knockout mutants. Nevertheless, the hypothesis of optimal growth is largely debatable for such genetically engineered mutants, as their metabolism was not exposed to evolutionary pressure. Basing on the assumption that metabolism in a knockout mutant operates as closely as possible to metabolism in the wild-type strain, two specific methods were introduced. They predict knockout mutant growth phenotypes by minimizing either the overall flux change [MoMA (Segrè *et al.*, 2002)] or the number of regulatory changes [ROOM (Shlomi *et al.*, 2005)] between the wild-type strain and the mutant strain (see Table 5). Both methods were shown to provide slightly better predictions than FBA.

The throughput of experiments evaluating qualitative growth phenotypes – i.e. described simply as *viable* or *lethal* – has increased dramatically in the last few years. Phenotype Microarrays from Biolog Inc. typically report growth phenotypes for several hundreds of media in a single experiment (Bochner *et al.*, 2001). In parallel to this, collections of knockout mutants are being built for a growing number of bacteria (Akerley *et al.*, 2002; Jacobs *et al.*, 2003; Kobayashi *et al.*, 2003; Baba *et al.*, 2006; Liberati *et al.*, 2006; Suzuki *et al.*, 2006; de Berardinis *et al.*, 2008). The systematic assessment of growth phenotypes of knockout mutants provides a significant resource for exploring the metabolic capabilities of organisms and investigating their gene functions (Carpenter & Sabatini, 2004), but their direct interpretation is made difficult by the complexity and size of metabolic networks (Gerdes *et al.*, 2006). These results can be readily compared with model predictions, however, providing a way to interpret them and assess the model correctness. Given the qualitative nature of these growth phenotypes, two types of inconsistencies may arise: *false viable* predictions – growth was predicted yet not observed experimentally – and *false lethal* predictions – growth was not predicted yet observed experimentally. On the one hand, these inconsistencies may be caused by limitations of the model or cases where the modeling assumptions do not hold. Regulation may for instance trigger a lethal phenotype by blocking an alternate pathway, which would not be predicted as blocked in the merely metabolic model. On the other hand, examining the inconsistencies may identify errors in the model and lead to its refinement. All model components may comprise errors, including the GPR associations, the metabolic network itself, and the stated biomass requirements. False lethal predictions are often clues that some biomass component is actually not essential, or that the model lacks an alternative gene or pathway that would allow it to survive in the given experimental conditions. Conversely, false viable predictions can help detect missing essential biomass components, genes falsely annotated as encoding isozymes or reactions that were wrongly assigned or are inactive in the experimental conditions (Duarte *et al.*, 2004; Joyce *et al.*, 2006). Growth phenotype predictions have been evaluated for a significant proportion of reconstructed models, whenever experimental data were available (see Table 4). Interpretation of inconsistent cases by expert examination led to several annotation and model refinements, some of which were supported by the results of targeted experiments (Covert *et al.*, 2004; Duarte *et al.*, 2004; Joyce *et al.*, 2006; Reed *et al.*, 2006b). Automated methods were recently introduced to systematically look for interpretations of inconsistencies and possible modifications in the model. Corrections of the GPR associations can be systematically proposed that match the gene essentiality observation with predicted reaction essentiality (M. Durot *et al.*, unpublished data). With regard to the metabolic network itself, metabolic gap filling approaches have been adapted to propose network corrections that resolve wrongly predicted growth phenotypes (Reed *et al.*, 2006b). Finally, valuable insights into the determination of essential biomass precursors can be provided by methods that analyze correlations between lethality and metabolite production (Imielinski *et al.*, 2005; Kim *et al.*, 2007). All these methods act independently on distinct components of the model. A unifying method integrating all types of corrections, which is yet to come, could lead to an integrated platform for the systematic interpretation of upcoming growth phenotyping results.

Models can actually predict growth phenotypes for any environmental condition and any combination of gene deletions, which is beyond reach of experiments. Given the combinatorial complexity of mixing several gene deletions, dedicated methods have been designed to analyze the effects of multiple deletions and applied to identify epistatic interactions between genes (Klamt & Gilles, 2004; Deutscher *et al.*,

Table 5. Main analytical methods for genome-scale models sorted by type of application

Analysis of network properties	
Flux sampling: random sampling of flux distribution among the set of possible metabolic states	Almaas *et al*. (2004), Reed & Palsson (2004), Wiback *et al*. (2004)
Flux variability analysis: examination of flux variability for each reaction	Mahadevan & Schilling (2003)
Metabolic pathway analysis, elementary modes/extreme pathways: comprehensive description of all independent metabolic modes achievable in the metabolic network	Schilling *et al*. (2000), Schuster *et al*. (2000), Klamt & Stelling (2003)
Flux coupling: identification of reaction pairs whose fluxes are coupled	Burgard *et al*. (2004)
Metabolite coupling/evaluation of conserved metabolite pools	Nikolaev *et al*. (2005), Becker *et al*. (2006), Imielinski *et al*. (2006)
Prediction and interpretation of bacterial growth phenotypes	
Producibility analysis of biomass precursors	Imielinski *et al*. (2005)
FBA: quantitative prediction of growth yield by maximization of growth rate given bounded nutrient input rates	Varma & Palsson (1994a, b)
MOMA: prediction of gene deletion mutant flux distribution by minimizing overall flux changes with wild type	Segrè *et al*. (2002)
ROOM: prediction of gene deletion mutant growth by minimizing regulatory changes with wild type	Shlomi *et al*. (2005)
Identification of multiple gene deletion essentialities	Klamt & Gilles (2004), Deutscher *et al*. (2006), Imielinski & Belta (2008)
Model-based interpretation of experimental data	
Metabolic flux measurements	
Metabolic Flux Analysis using labeled metabolites: prediction of attainable reaction fluxes given observed metabolite isotopic patterns	Wiechert (2001), Sauer (2006)
Global prediction of reaction activities using metabolic flux measurements on subsets of reactions	Herrgård *et al*. (2006a, b)
Identification of metabolic objectives best describing observed fluxes	Burgard & Maranas (2003), Schuetz *et al*. (2007)
Metabolite concentrations	
Comparison of model coverage with experimentally detected metabolites	Oh *et al*. (2007)
NET analysis and TMFA: application of thermodynamic constraints to reaction directions using metabolite concentrations	Kümmel *et al*. (2006a, b), Henry *et al*. (2007)
Gene expression	
Identification of metabolic pathways correlated with gene expression levels	Schwartz *et al*. (2007)
Refinement of flux distribution predictions by blocking reactions corresponding to unexpressed genes	Akesson *et al*. (2004)
Evaluation of consistency of gene expression levels with metabolic objectives	Becker & Palsson (2008)
rFBA and SR-FBA: prediction of gene expression states using Boolean regulatory rules	Covert *et al*. (2001), Barrett *et al*. (2005), Barrett & Palsson (2006), Shlomi *et al*. (2007a, b)
Metabolic engineering	
Systematic identification of gene deletions enhancing metabolite production yield	Burgard *et al*. (2003), Patil *et al*. (2004), Alper *et al*. (2005a, b)
OptStrain: systematic identification of reaction additions enabling the production of novel metabolites	Pharkya *et al*. (2004)
Prediction of adjustments of enzyme expression levels enhancing metabolite production yield	Pharkya & Maranas (2006), Lee *et al*. (2007)

2006, 2008; Imielinski & Belta, 2008). Prediction of growth phenotypes have also been used to automatically assign condition-dependent roles to genes (Shlomi *et al*., 2007b), investigate the causes of gene dispensability (Papp *et al*., 2004; Kuepfer *et al*., 2005), or study bacterial evolution (Pál *et al*., 2005a, 2006). These two latter studies on bacterial evolution used an *E. coli* model to analyze the effect of changing growth environments on the acquisitionof new metabolic capabilities by horizontal gene transfer (Pál *et al*., 2005a) and to simulate the reductive evolution of metabolism in specific environmental conditions (Pál *et al*., 2006).

Model-based interpretation of experimental data

The recent development of experimental techniques has enabled measurements at genome-scale of several types of quantities, generating so-called 'omics' datasets. These datasets provide partial yet comprehensive snapshots of cellular mechanisms (Ishii *et al*., 2007a), but their interpretation is made difficult by the volume of data. Computational methods are thus needed if meaningful biological results are to be extracted (Joyce & Palsson, 2006). A variety of

methods have been developed to exploit experimental data related to metabolic states, for example measurements of metabolic fluxes, metabolite concentrations, enzyme levels, or gene expression, in the light of genome-scale models. Two cases generally arise: either experimental observations are directly comparable to model predictions, or these observations lead to the imposition of additional constraints that refine the set of predicted metabolic states. Observations falling in the second category allow for instance the selection of those metabolic routes that are compatible with the experimental observations, or help predict quantitative values for the fluxes. When directly comparable to model predictions, experimental data may be used to assess model correctness and assumptions, identify inconsistencies, and target improvements, as illustrated above with growth phenotypes (Reed *et al.*, 2006b). We will review such integration methods in the following sections for three types of experimental data: measurement of (1) reaction fluxes, (2) metabolite concentrations, and (3) gene expression levels.

Refining the model with experimental data increases its correctness with respect to the observations but may decrease its predictive power. Predictions performed with a refined model should actually be interpreted with care to avoid circular reasoning: data that have been directly used to improve the model can no more be considered as predictions, they are part of the evidences on which the model is based to perform predictions. For instance, a model whose maintenance parameters have been determined using growth rate measurement can no more *predict* the growth rate for the environmental condition. This problem can become serious when models are extensively fitted with experimental data, as they then become more descriptive than predictive. Nevertheless, some refinement processes applied to genome-scale models involve finding additional biological evidence that supports the refinement, thereby breaking the circular reasoning. For instance, corrections of inconsistent growth phenotype predictions by additions of alternate enzymes often involve finding additional proofs that the introduced enzymes possess the right activity.

Metabolic flux measurements

One of the most direct experimental accesses to metabolic fluxes is provided by atom-labeling experiments (Wiechert, 2001; Sauer, 2006). By analyzing the fate of labeled metabolites, valuable information can be deduced about the reactions that are actually taking place. The most common technique for this consists in analyzing the stable isotope patterns (mostly using ^{13}C) found in products of metabolism given known isotope patterns in nutrient metabolites (Wiechert, 2001; Sauer, 2006). These data can be properly interpreted only using a metabolic model that includes information about atom mappings for each reaction (Zupke & Stephanopoulos, 1994; Wiechert *et al.*, 1999; Antoniewicz *et al.*, 2007a). Such models have been built for a few organisms, often using existing constraint-based models as a basis (Antoniewicz *et al.*, 2007b; Suthers *et al.*, 2007). While atom mappings for reactions are currently mostly inferred using chemoinformatics methods (Raymond *et al.*, 2002; Arita, 2003; Hattori *et al.*, 2003), this information will likely be made accessible in dedicated databases in the coming years.

By qualitatively examining isotope patterns in nutrients and products, information can already be extracted about the possible routes of conversion (van Winden *et al.*, 2001; Sauer, 2006; Kuchel & Philp, 2008). Patterns in products actually depend on their biosynthetic pathways. Observed patterns that are inconsistent with the predicted possible patterns are clues that other pathways may occur *in vivo*. This approach was for instance recently used to evaluate the model of *G. sulfurreducens*: an inconsistent isotope pattern for isoleucine led to the discovery of an isoleucine biosynthesis pathway previously uncharacterized in this bacteria (Risso *et al.*, 2008).

Quantitative interpretation of isotope patterns together with measurement of extracellular metabolite fluxes can help determine the value of intracellular reaction fluxes using Metabolic Flux Analysis (Zupke & Stephanopoulos, 1994; Stephanopoulos *et al.*, 1998; Wiechert *et al.*, 1999; Sauer, 2006; Antoniewicz *et al.*, 2007a). Known flux values can then be directly exploited in models to characterize which metabolic pathways are operating and quantify their fluxes. As an application, Herrgård *et al.* (2006a) introduced the optimal metabolic network identification method, which combines flux measurements for a fraction of the reactions with the assumption of optimal growth from FBA to globally infer which reactions are active. This method has been for instance used to identify bottleneck reactions that limit the growth in engineered strains, and discard putative reactions from newly reconstructed models (Herrgård *et al.*, 2006a).

Observed fluxes were also used to determine relevant objective functions to choose when predicting metabolic states with FBA (Burgard & Maranas, 2003). By evaluating the match of predicted fluxes with observed ones, these studies could identify those metabolic objectives that provided the best fit. Distinct objectives, including maximization of ATP or biomass yields, were identified for instance in *E. coli* depending on the environmental conditions (Schuetz *et al.*, 2007). Observed metabolic fluxes, however, often show that metabolism does not necessarily operate according to optimality principles (Fischer *et al.*, 2004), especially when regulatory constraints are overlooked.

Metabolite concentrations

High-throughput measurement of intracellular metabolite concentrations is becoming common practice thanks to

recent developments in MS and NMR technologies (Dunn *et al.*, 2005; Dettmer *et al.*, 2007). Metabolite profiling experiments commonly detect thousands of peaks, among which hundreds can usually be exploited to identify metabolites and determine their concentrations, using for instance known spectra of reference metabolites (Dunn *et al.*, 2005). These datasets, while not fully comprehensive, provide significant information on metabolites present in the cell.

Merely comparing the set of detected metabolites to the set of metabolites present in the model already help in assessing the comprehensiveness of the model. For example, in the reconstruction process of *Bacillus subtilis* metabolic model, Oh *et al.* (2007) evaluated the overlap between model metabolites and intracellular metabolites identified in a metabolomics dataset; among 350 intracellular metabolites identified, only 160 were present in the model. No previously known biochemical activities could be associated with the remaining metabolites, illustrating the fact that a large part of *B. subtilis* metabolism remains unknown. These unaccounted metabolites can guide further investigations on missing activities, leading to expansion of the model's metabolite scope consequently.

By extending the constraint-based modeling framework to encompass thermodynamic constraints on Gibbs energies of reactions, knowledge of absolute metabolite concentrations can be translated into constraints on flux directions (Kümmel *et al.*, 2006b; Henry *et al.*, 2007). A first application is to check the consistency of metabolomic datasets with respect to metabolic fluxes predicted by the model. Methods and software have been developed to pinpoint inconsistent concentration measures (Zamboni *et al.*, 2008). Conversely, metabolomic-derived constraints refine the characterization of metabolic fluxes within the model; their integration has allowed the prediction of ranges of concentrations for unmeasured metabolites, reaction directions, and ranges of Gibbs energies of reactions, identifying thereby potentially regulated reactions (Kümmel *et al.*, 2006b).

Thermodynamic constraints merely enforce link between the concentrations of metabolites and the directions of reactions. Taking reaction kinetics into consideration could reinforce that link and make it more quantitative. Extending models to handle kinetics is still an open issue (Famili *et al.*, 2005; Yugi *et al.*, 2005; Ishii *et al.*, 2007b; Smallbone *et al.*, 2007; Covert *et al.*, 2008; Jamshidi & Palsson, 2008), all the more challenging because of the potential influence of regulation, the scarcity of kinetic parameter values and the lack of scalable analytical methods.

Gene expression data

Thanks to technological advances, gene expression levels are among the most widely accessible type of 'large-scale' experimental data. While such datasets provide a global overview of the level of expression of enzymes, deriving information on reaction fluxes from gene expression levels is hindered by the numerous biological processes intervening between them. Changes in rates of translation or mRNA and enzyme degradation may significantly modify the quantity of enzymes available from a given amount of transcript. In addition, changes in substrate/product concentrations or metabolic regulations can influence the reaction fluxes irrespective to the enzyme quantities. As a consequence, no simple correlations are necessarily observed between gene expression levels and reaction fluxes (Gygi *et al.*, 1999; ter Kuile & Westerhoff, 2001; Yang *et al.*, 2002; Akesson *et al.*, 2004).

Some approaches have nonetheless been developed to exploit information from gene expression data using models. In the vein of pathway- or module-based methods interpreting changes of gene expressions at the level of pathways or biological processes (Hanisch *et al.*, 2002; Draghici *et al.*, 2003; Yang *et al.*, 2004), methods relying on a graph representation of metabolism (Patil & Nielsen, 2005) or on a decomposition of metabolic models into elementary modes (Schwartz *et al.*, 2007) were introduced to correlate expression levels with possible metabolic states. These approaches are merely descriptive: the model provides a suitable metabolic context to interpret the experimental data. Gene expression data have also been used to refine the characterization of metabolic fluxes in models. For instance, by blocking reactions corresponding to unexpressed genes, metabolic fluxes could be characterized more precisely in a yeast model (Akesson *et al.*, 2004). In the same spirit, a method was recently introduced to evaluate the consistency of gene expression datasets with metabolic objectives, and identify subsets of active reactions that best correlate with expressed genes and metabolic objectives (Becker & Palsson, 2008). Even though these methods only rely on a limited dependency between gene expression level and reaction flux – reactions catalyzed by unexpressed genes should have low fluxes – they succeed in somewhat improving the characterization of metabolic states, or in assessing the consistency of the model with the experimental data.

As an attempt to account for transcriptional regulation, regulatory interactions were introduced in models by translating them into Boolean rules (Covert *et al.*, 2001). In such joint regulatory-metabolic model, Boolean variables qualitatively describe the transcription state of genes, including genes coding for enzymes and transcription factors, while Boolean rules determine their regulatory dependencies. Metabolic reactions are then allowed to have a nonzero flux only if the transcriptional state of their enzymes is *true*. Several methods have been developed to study these joint models. Regulatory FBA (rFBA) simulates time courses of gene expression states: at each time step, the new transcriptional state is computed from the metabolic state predicted at the previous time step, and is used to constrain FBA

prediction of the current metabolic state (Covert *et al.*, 2001). A specific representation scheme was later developed to encode the sequence of expression states predicted by rFBA in a unified manner, in order to compare regulatory responses across various environments (Barrett *et al.*, 2005). Another type of method has been recently developed to determine joint steady states of gene expression and metabolic fluxes. Examining these steady states contributed to the identification of redundantly expressed enzymes and the quantification of the effect of transcriptional regulation in determining flux activity in *E. coli* (Shlomi *et al.*, 2007a). Finally, two studies compared experimental expression levels with predicted expression states to assess the correctness of joint regulatory-metabolic models of *E. coli* and yeast (Covert *et al.*, 2004; Herrgård *et al.*, 2006b). A significant proportion of inconsistent expression states could be corrected in these models by searching for missing interactions (Covert *et al.*, 2004; Herrgård *et al.*, 2006b). In the same vein, a method was recently designed to automate the identification of experiments that are likely to bring most information on potentially missing regulatory interactions (Barrett & Palsson, 2006).

Using genome-scale models for metabolic engineering

The use of microbial organisms for industrial purposes has grown considerably in the past few years, with potential applications ranging from the production of valuable metabolites to the degradation of pollutants and the generation of renewable energy (Janssen *et al.*, 2005; Ro *et al.*, 2006; Peng *et al.*, 2008; Rittmann, 2008). The field of metabolic engineering aims at designing and improving industrial microorganisms through the rational design of genetic manipulations leading to enhanced performance (Bailey, 1991; Stephanopoulos *et al.*, 1998). With the advent of genome-scale experimental technologies, the set of metabolic engineering methods is progressively expanding to include systems-wide analyses, enabling for instance to study the operation of regulatory and metabolic networks at large scale (Park *et al.*, 2008). In this respect, genome-scale metabolic models provide to engineers an effective toolbox to investigate the metabolic behavior of their strain of interest and target improvements (Kim *et al.*, 2008).

As a first class of applications, all analytical methods presented in the previous sections can be directly applied to engineering purposes. Such methods may help for instance to evaluate the maximum theoretical efficiencies of pathways or determine appropriate host strains by predicting their metabolic capabilities from their reconstructed models. More importantly, metabolic models can help in characterizing the actual metabolism operation of engineered strains, especially when experimental data have been acquired on them. Metabolic Flux Analysis provides for instance quantitative values for intracellular fluxes, which may be used to determine the actual pathway utilization and pinpoint bottleneck reactions (Stephanopoulos *et al.*, 1998). Such information is of high significance for the metabolic engineers, as it may help them in designing further metabolic modifications.

Metabolic models also provide the ability to formulate hypotheses and evaluate *in silico* the potential of genetic modifications. A common cause of low production yields lies in the presence of pathways that divert fluxes to the production of undesirable byproducts or compete for the utilization of precursors and cofactors. While such pathways may be identified manually, their direct removal through gene deletion may cause side effects, for example alter the regeneration of cofactors, the redox balance, or the energy balance (Kim *et al.*, 2008). Genome-scale models can predict the effect of gene deletions on metabolic phenotypes. Several methods were designed with the aim of selecting those gene deletions that would provide the greatest benefit for a given metabolite production goal. Alper *et al.* (2005a) developed a procedure that sequentially screen the effect of single and multiple gene deletions in order to select those enabling the best product yields while maintaining sufficient growth rates. They successfully applied their method to enhance the yield of a lycopene producing *E. coli* strain (Alper *et al.*, 2005b). Screening *in silico* the high number of combinations of multiple gene deletions may turn out to be costly and practically impossible. Optimization methods based on genetic (Patil *et al.*, 2005) or linear programming (Burgard *et al.*, 2003) algorithms were introduced to circumvent this issue. The second optimization method, called OptKnock, specifically searches gene deletions coupling the production of a targeted metabolite with growth rate; the rationale being that improving the growth rate by adaptive evolution would jointly improve the metabolite production rate and that this coupling would make the engineered strain more evolutionary stable (Burgard *et al.*, 2003). Gene deletions proposed by this method were tested experimentally to enhance lactic acid production in an *E. coli* strain (Fong *et al.*, 2005). Adaptive evolution experiments performed on the engineered strains actually showed that lactic acid production was coupled to growth and achieved increased secretion rates of the product. In addition to gene deletions, metabolic models can explore the effect of adding new pathways, and help select the most appropriate ones. In this aim, the OptStrain method was designed to systematically suggest additions of reactions to produce novel metabolites (Pharkya *et al.*, 2004). OptStrain relies on a comprehensive database of biochemical reactions and may propose alternative solutions. A last set of methods consists in designing suitable up- or downregulations of metabolic enzymes. Intervening on gene expression levels is indeed a powerful

tool to tune metabolism operation, but the specific effects of such interventions are often hardly predictable (Kim *et al.*, 2008). In a study involving a L-threonine producing strain of *E. coli*, Lee *et al.* (2007) made use of its metabolic model to predict gene expression changes enhancing the strain yield. Specifically, they predicted flux values of key reactions leading to optimal L-threonine production and compared them with measured fluxes. They then used the relative difference between them to guide the tuning of the expression of the corresponding genes. A more systematic approach was introduced with the OptReg method, which identifies at genome-scale the relative changes of flux values with respect to the wild-type flux distribution that provide the best production yield (Pharkya & Maranas, 2006). Results of OptReg can be used to identify candidate enzymes for up- or downregulation.

Yet, two main issues limit the predictive capabilities of metabolic models. First, while regulation may play a central role in controlling the efficiency of product synthesis, it is completely overlooked in metabolic models. Studying regulatory interactions – using for instance models of regulatory networks – may actually provide useful insights, for example to remove feedback inhibitions or fine-tune transcriptional regulatory circuits commanding the product biosynthesis (Kim *et al.*, 2008). Not accounting for enzyme quantities but only reaction fluxes imposes a second limitation to genome-scale models. Implementing changes in flux values – suggested for instance by metabolic model optimization methods – by altering the quantity of enzymes is a difficult task, as enzyme kinetics and metabolite concentrations may significantly influence the flux change. In order to determine the effect of enzyme quantity changes on metabolic fluxes, more detailed approaches are required, for example metabolic control analysis (Fell, 1992).

Resources, databases, and tools

At the time of this review, genome-scale models have been reconstructed for at least 17 bacteria (see Table 4). For all of them, extensive manual curation was required in order to integrate information from the literature on their biochemistry and physiology with functional information from genome annotation. These models are therefore of high quality on average, and mostly complete with respect to the current knowledge of their metabolism. An increasing subset is being assessed and corrected against large-scale experimental data (see Table 4), and an impressive array of analytical studies has been applied to the most popular ones, for example *E. coli* (Feist & Palsson, 2008).

Models used to be made available independently by their authors, under a variety of naming conventions and formats. This is a significant obstacle to their reusability, as significant effort is required to adapt them to modeling software other than the ones they were constructed with. Differences in reaction and metabolite names also hamper direct comparisons between different models. Fortunately, some attempts to address these issues are under way. The general-purpose SBML format (Systems Biology Markup Language) (Hucka *et al.*, 2003) is often used to exchange constraint-based models, thus playing the role of a 'default' standard for models. While SBML can be imported by many modeling tools, it is not fully adapted to the specifics of models; this may result in information or functionality loss during exchange. In addition to providing a standard format, SBML supports the association of model components with external references, such as reaction and metabolite identifiers in universal metabolic databases, using MIRIAM annotations (Le Novère *et al.*, 2005). If widely used, this feature should facilitate model reuse and comparison.

In order to facilitate model reuse and comparison, dedicated model repositories have been developed. Perhaps the most widely adopted initiative of this type is the Biomodels.net repository (Le Novère *et al.*, 2006) which stores biochemical models of any type in SBML format. Because of its focus on more detailed dynamic models and the related generic format choice, the repository is not fully compatible with constraint-based models and qualitative predictions, as illustrated by the current low number of such models included. Agreements with several journals make it mandatory for authors to deposit models mentioned in their manuscripts in Biomodels.net, where they are checked for syntactic correctness. On some models, a more elaborate test on the compatibility between model predictions and results presented in the associated paper is also performed.

Currently, the only freely accessible (to academic users) repository dedicated to constraint-based models is the BiGG database (http://bigg.ucsd.edu). Its unified dictionary of metabolite and reaction names enables direct comparisons between its metabolic models.

Relatively few software tools have been specifically developed to handle genome-scale constraint-based models, compared with the number of tools developed for kinetic modeling. As the modeling framework relies primarily on linear algebra and linear programming, general purpose mathematical software platforms, for example MATLAB (http://www.mathworks.com/) and MATHEMATICA (http://www.wolfram.com/), or optimization modeling packages, for example GAMS (http://www.gams.com/), are well suited. Specialized optimization packages can be added for greater efficiency. In addition, modules dedicated to constraint-based modeling have been developed for MATLAB: FLUXANALYZER (Klamt *et al.*, 2007), the COBRA TOOLBOX (Becker *et al.*, 2007), or METATOOL (von Kamp & Schuster, 2006) for elementary mode analysis are good representatives. Libraries for importing SBML models within these programs are also provided by the SBML developer community (Bornstein

et al., 2008). Among the software tools that are stand-alone, one should mention the SYSTEMS BIOLOGY RESEARCH TOOLBOX (Wright & Wagner, 2008), SCRUMPY (Poolman, 2006), METAFLUXNET (Lee *et al.*, 2003), or FLUXEXPLORER (Luo *et al.*, 2006), each with their own specific strengths. Interestingly, very few programs focus or even support the model reconstruction process by providing the analytical capabilities for consistency checks: the commercial SYMPHENY platform (http://www.genomatica.com/) associates a metabolic database with several analytical methods, while YANASQUARE (Schwarz *et al.*, 2007) facilitates the reconstruction of models from KEGG and performs selected structural analyses (e.g. elementary modes). Very recently, web-based tools have been released to enable on-line analyses on specific metabolic models (Beste *et al.*, 2007; Durot *et al.*, 2008). Given the need for faster and better reconstruction, we expect more progress in that direction.

Concluding remarks and future directions

Constraint-based genome-scale metabolic models can be viewed as 'systems-level' analytical layers which enable computation and reasoning on the consequences of the accumulated knowledge on the biochemistry encoded in a given genome, and confrontation of that knowledge with the known physiology of the corresponding species or with additional experimental evidence. These models thus bridge the gap between genotype and phenotype and enable a wide spectrum of analyses and *in silico* experiments, providing a solid foundation for systems analyses and metabolic engineering.

The systematic and automated reconstruction of genome-scale models from genomes and additional high-throughput data may seem like a natural extension of genome annotation (Reed *et al.*, 2006a), but remains beyond the reach of current methods. While genome-scale models can be reconstructed using only sequence and qualitative functional information, gaining the additional predictive and analytical power of models still requires significant effort and expertise. Genome annotations must first be translated into a network, which must then be turned into a model with the help of additional information, and systematically checked with respect to biochemical consistency rules and experimental observations. Only after a model is complete enough to enable meaningful predictions at the phenotypic level can it be used to predict phenotypes or other properties beyond those that can be immediately verified.

Obstacles to automating this process include technical difficulties in translating annotations into proper biochemical activities, and also the fact that methods for model refinement have been designed and applied separately for each type of experimental data. There is increasing pressure for this situation to evolve, however, as the boost in the throughput of experimental techniques and the advent of 'multi-omics' datasets (Ishii *et al.*, 2007a) promises a wealth of information that will be exploitable only by computer-assisted interpretation, with the help of models. At the same time, the field of metabolic modeling is now approaching the level of maturity necessary for several data integration methods to be used together as components in integrated model reconstruction and refinement strategies.

Significant benefits could result from the availability of a wider spectrum of bacterial metabolic models. They would provide an integrated view of metabolic pathways across the tree of life, thereby enabling so-called transverse approaches to annotation, and a variety of comparative metabolic analysis. To that end, the notion of pathway – defined unambiguously as the conversion between specified sets of input compounds (reactants) and output compounds (products) – can bring a useful decomposition of metabolism into basic biochemical functional units, in the spirit pioneered by SEED (Overbeek *et al.*, 2005), KEGG Modules (Kanehisa *et al.*, 2007), or MetaCyc (Caspi *et al.*, 2006). The field of bacterial evolution is poised to benefit as well: for instance, the availability of models for several bacteria along the phylogenetic tree would allow more comprehensive studies on the constraints implied by bacteria's metabolic capabilities and their evolution. While this type of study has been pioneered with a few selected models (Pál *et al.*, 2005a, 2006), working with a larger set of models will undoubtedly bring different insights (see (Kreimer *et al.*, 2008) for an example with networks). Modeling can also help in studying bacterial communities, as chemical interactions occurring between bacteria often need to be understood within the context of their metabolisms. Indeed, models have already been reconstructed and analyzed for small communities (Stolyar *et al.*, 2007); progress on that front may prove very useful in studying metabolic interactions in more complex communities, assisting in the functional interpretation of metagenome sequences. Last but not least, metabolic engineering applications would clearly benefit from the availability of a large set of bacterial models, as these would constitute a repository of characterized metabolic pathways, facilitating the combinatorial design of new catalytic systems, providing solid bases to test hypothetical genetic constructions, and helping with the selection of relevant strains for specific engineering objectives.

Acknowledgements

We would like to thank the two anonymous reviewers for their numerous suggestions, which helped improve the manuscript. We are grateful for the support of the European Networks of Excellence BIOSAPIENS (contract no. LSHG-CT-2003-503265) and ENFIN (contract no. LSHG-CT-2005-518254).

Statement

References

Aghaie A, Lechaplais C, Sirven P *et al.* (2008) New insights into the alternative D-glucarate degradation pathway. *J Biol Chem* **283**: 15638–15646.

Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N & Mekalanos JJ (2002) A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *P Natl Acad Sci USA* **99**: 966–971.

Akesson M, Förster J & Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metab Eng* **6**: 285–293.

Almaas E, Kovacs B, Vicsek T, Oltvai ZN & Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**: 839–843.

Alper H, Jin Y-S, Moxley JF & Stephanopoulos G (2005a) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* **7**: 155–164.

Alper H, Miyaoku K & Stephanopoulos G (2005b) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* **23**: 612–616.

Antoniewicz MR, Kelleher JK & Stephanopoulos G (2007a) Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab Eng* **9**: 68–86.

Antoniewicz MR, Kraynie DF, Laffend LA, González-Lergier J, Kelleher JK & Stephanopoulos G (2007b) Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metab Eng* **9**: 277–292.

Apweiler R, Attwood TK, Bairoch A *et al.* (2000) InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.

Arakawa K, Yamada Y, Shinoda K, Nakayama Y & Tomita M (2006) GEM system: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* **7**: 168.

Arita M (2000) Metabolic reconstruction using shortest paths. *Simulat Pract Theory* **8**: 109–125.

Arita M (2003) *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Res* **13**: 2455–2466.

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* **25**: 25–29.

Baart G, Zomer B, de Haan A, van der Pol L, Beuvery EC, Tramper J & Martens D (2007) Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes. *Genome Biol* **8**: R136.

Baba T, Ara T, Hasegawa M *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**: 2006.0008.

Bailey JE (1991) Toward a science of metabolic engineering. *Science* **252**: 1668–1675.

Barrett CL & Palsson BO (2006) Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS Comput Biol* **2**: e52.

Barrett CL, Herring CD, Reed JL & Palsson BO (2005) The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *P Natl Acad Sci USA* **102**: 19103–19108.

Barthelmes J, Ebeling C, Chang A, Schomburg I & Schomburg D (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* **35**: D511–D514.

Beard DA, Liang S-D & Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophys J* **83**: 79–86.

Beard DA, Babson E, Curtis E & Qian H (2004) Thermodynamic constraints for biochemical networks. *J Theor Biol* **228**: 327–333.

Becker SA & Palsson BO (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol* **5**: 8.

Becker SA & Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol* **4**: e1000082.

Becker SA, Price ND & Palsson BØ (2006) Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics* **7**: 111.

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ & Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat Protoc* **2**: 727–738.

Besemer J, Lomsadze A & Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607–2618.

Beste D, Hooper T, Stewart G *et al.* (2007) GSMN-TB: a web-based genome scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biol* **8**: R89.

Bochner BR, Gadzinski P & Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* **11**: 1246–1255.

Bocs S, Cruveiller S, Vallenet D, Nuel G & Médigue C (2003) AMIGene: Annotation of microbial genes. *Nucleic Acids Res* **31**: 3723–3726.

Bornstein BJ, Keating SM, Jouraku A & Hucka M (2008) LibSBML: an API library for SBML. *Bioinformatics* **24**: 880–881.

Borodina I, Krabben P & Nielsen J (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res* **15**: 820–829.

Boutet E, Lieberherr D, Tognolli M, Schneider M & Bairoch A (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol Biol* **406**: 89–112.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO & Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**: R35.

Boyer F & Viari A (2003) *Ab initio* reconstruction of metabolic pathways. *Bioinformatics* **19**(suppl 2): ii26–ii34.

Braunstein A, Mulet R & Pagnani A (2008) Estimating the size of the solution space of metabolic networks. *BMC Bioinformatics* **9**: 240.

Breitling R, Ritchie S, Goodenowe D, Stewart ML & Barrett MP (2006) *Ab initio* prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* **2**: 155–164.

Burgard AP & Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng* **82**: 670–677.

Burgard AP, Pharkya P & Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* **84**: 647–657.

Burgard AP, Nikolaev EV, Schilling CH & Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14**: 301–312.

Carpenter AE & Sabatini DM (2004) Systematic genome-wide screens of gene function. *Nat Rev Genet* **5**: 11–22.

Caspi R, Foerster H, Fulcher CA *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **34**: D511–D516.

Chen L & Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* **7**: R17.

Claudel-Renard C, Chevalet C, Faraut T & Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**: 6633–6639.

Covert MW, Schilling CH & Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213**: 73–88.

Covert MW, Knight EM, Reed JL, Herrgard MJ & Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**: 92–96.

Covert MW, Xiao N, Chen TJ & Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24**: 2044–2050.

de Berardinis V, Vallenet D, Castelli V *et al.* (2008) A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol Syst Biol* **4**: 174.

DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M & Best A (2007) Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* **8**: 139.

Delcher AL, Harmon D, Kasif S, White O & Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–4641.

Dettmer K, Aronov PA & Hammock BD (2007) Mass spectrometry-based metabolomics. *Mass Spectrom Rev* **26**: 51–78.

Deutscher D, Meilijson I, Kupiec M & Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* **38**: 993–998.

Deutscher D, Meilijson I, Schuster S & Ruppin E (2008) Can single knockouts accurately single out gene functions? *BMC Syst Biol* **2**: 50.

Draghici S, Khatri P, Martins RP, Ostermeier GC & Krawetz SA (2003) Global functional profiling of gene expression. *Genomics* **81**: 98–104.

Duarte NC, Herrgard MJ & Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14**: 1298–1309.

Dunn WB, Bailey NJC & Johnson HE (2005) Measuring the metabolome: current analytical technologies. *Analyst* **130**: 606–625.

Durot M, Le Fevre F, de Berardinis V *et al.* (2008) Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol* **2**: 85.

Ebenhöh O, Handorf T & Heinrich R (2004) Structural analysis of expanding metabolic networks. *Genome Infor* **15**: 35–45.

Edwards JS & Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *P Natl Acad Sci USA* **97**: 5528–5533.

Edwards JS, Ibarra RU & Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19**: 125–130.

Edwards JS, Ramakrishna R & Palsson BO (2002) Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng* **77**: 27–36.

Ellis LBM, Roe D & Wackett LP (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res* **34**: D517–D521.

Ellis LBM, Gao J, Fenner K & Wackett LP (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res* **36**: W427–W432.

Famili I, Mahadevan R & Palsson BO (2005) k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J* **88**: 1616–1625.

Feist AM & Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26**: 659–667.

Feist AM, Scholten JCM, Palsson BØ, Brockman FJ & Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol Syst Biol* **2**: 2006.0004.

Feist AM, Henry CS, Reed JL *et al.* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3**: 121.

Fell DA (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J* **286**: 313–330.

Fischer E, Zamboni N & Sauer U (2004) High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived ^{13}C constraints. *Anal Biochem* **325**: 308–316.

Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD & Palsson BO (2005) *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* **91**: 643–648.

Fuhrer T, Chen L, Sauer U & Vitkup D (2007) Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. *J Bacteriol* **189**: 8073–8078.

Gasteiger J (2005) Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem* **384**: 57–64.

Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R & Osterman A (2006) Essential genes on metabolic maps. *Curr Opin Biotech* **17**: 448–456.

Gevorgyan A, Poolman MG & Fell DA (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* **24**: 2245–2251.

Green ML & Karp PD (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**: 76.

Gygi SP, Rochon Y, Franza BR & Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**: 1720–1730.

Hanisch D, Zien A, Zimmer R & Lengauer T (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics* **18**(suppl 1): S145–S154.

Hattori M, Okuno Y, Goto S & Kanehisa M (2003) Heuristics for chemical compound matching. *Genome Infor* **14**: 144–153.

Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD & Broadbelt LJ (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**: 1603–1609.

Heinemann M, Kümmel A, Ruinatscha R & Panke S (2005) *In silico* genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. *Biotechnol Bioeng* **92**: 850–864.

Henry CS, Broadbelt LJ & Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. *Biophys J* **92**: 1792–1805.

Herrgård MJ, Fong SS & Palsson BØ (2006a) Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* **2**: e72.

Herrgård MJ, Lee B-S, Portnoy V & Palsson BØ (2006b) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res* **16**: 627–635.

Hong SH, Kim JS, Lee SY *et al.* (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol* **22**: 1275–1281.

Huang M, Oppermann-Sanio FB & Steinbüchel A (1999) Biochemical and molecular characterization of the *Bacillus subtilis* acetoin catabolic pathway. *J Bacteriol* **181**: 3837–3841.

Hucka M, Finney A, Sauro HM *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531.

Ibarra RU, Edwards JS & Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**: 186–189.

Imielinski M & Belta C (2008) Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst Biol* **2**: 40.

Imielinski M, Belta C, Halasz A & Rubin H (2005) Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities. *Bioinformatics* **21**: 2008–2016.

Imielinski M, Belta C, Rubin H & Halász A (2006) Systematic analysis of conservation relations in *Escherichia coli* genome-scale metabolic network reveals novel growth media. *Biophys J* **90**: 2659–2672.

Ishii N, Nakahigashi K, Baba T *et al.* (2007a) Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* **316**: 593–597.

Ishii N, Nakayama Y & Tomita M (2007b) Distinguishing enzymes using metabolome data for the hybrid dynamic/static method. *Theor Biol Med Model* **4**: 19.

Jacobs MA, Alwood A, Thaipisuttikul I *et al.* (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *P Natl Acad Sci USA* **100**: 14339–14344.

Jamshidi N & Palsson B (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol* **1**: 26.

Jamshidi N & Palsson BØ (2008) Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol* **4**: 171.

Janssen DB, Dinkla IJT, Poelarends GJ & Terpstra P (2005) Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol* **7**: 1868–1882.

Joyce AR & Palsson BØ (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Bio* **7**: 198–210.

Joyce AR, Reed JL, White A *et al.* (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J Bacteriol* **188**: 8259–8271.

Kanehisa M, Araki M, Goto S *et al.* (2007) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**: D480–D484.

Karp PD, Paley S & Romero P (2002) The pathway tools software. *Bioinformatics* **18**(suppl 1): S225–S232.

Karp PD, Keseler IM, Shearer A *et al.* (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res* **35**: 7577–7590.

Kharchenko P, Chen L, Freund Y, Vitkup D & Church GM (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* **7**: 177.

Kim HU, Kim TY & Lee SY (2008) Metabolic flux analysis and metabolic engineering of microorganisms. *Mol Biosyst* **4**: 113–120.

Kim P-J, Lee D-Y, Kim TY, Lee KH, Jeong H, Lee SY & Park S (2007) Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *P Natl Acad Sci USA* **104**: 13638–13642.

Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H & Mori H (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res* **12**: 291–299.

Klamt S & Gilles ED (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics* **20**: 226–234.

Klamt S & Stelling J (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol* **21**: 64–69.

Klamt S, Saez-Rodriguez J & Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol* **1**: 2.

Klopman G, Dimayuga M & Talafous J (1994) META. 1. A program for the evaluation of metabolic transformation of chemicals. *J Chem Inf Comput Sci* **34**: 1320–1325.

Kobayashi K, Ehrlich SD, Albertini A *et al.* (2003) Essential *Bacillus subtilis* genes. *P Natl Acad Sci USA* **100**: 4678–4683.

Kreimer A, Borenstein E, Gophna U & Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. *P Natl Acad Sci USA* **105**: 6976–6981.

Kuchel PW & Philp DJ (2008) Isotopomer subspaces as indicators of metabolic-pathway structure. *J Theor Biol* **252**: 391–401.

Kuepfer L, Sauer U & Blank LM (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* **15**: 1421–1430.

Kumar VS, Dasika MS & Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**: 212.

Kümmel A, Panke S & Heinemann M (2006a) Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* **7**: 512.

Kümmel A, Panke S & Heinemann M (2006b) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* **2**: 2006.0034.

Lee J, Yun H, Feist A, Palsson B & Lee S (2008a) Genome-scale reconstruction and *in silico* analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl Microbiol Biot* **80**: 849–862.

Lee JM, Gianchandani EP, Eddy JA & Papin JA (2008b) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol* **4**: e1000086.

Lee KH, Park JH, Kim TY, Kim HU & Lee SY (2007) Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* **3**: 149.

Lee SY, Lee D-Y, Hong SH, Kim TY, Yun H, Oh Y-G & Park S (2003) MetaFluxNet, a program package for metabolic pathway construction and analysis, and its use in large-scale metabolic flux analysis of *Escherichia coli*. *Genome Infor* **14**: 23–33.

Le Novère N, Finney A, Hucka M *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**: 1509–1515.

Le Novère N, Bornstein B, Broicher A *et al.* (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34**: D689–D691.

Liberati NT, Urbach JM, Miyata S *et al.* (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *P Natl Acad Sci USA* **103**: 2833–2838.

Luo R, Liao S, Zeng S, Li Y & Luo Q (2006) FluxExplorer: a general platform for modeling and analyses of metabolic networks based on stoichiometry. *Chin Sci Bull* **51**: 689–696.

Ma H & Zeng A-P (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**: 270–277.

Mahadevan R & Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**: 264–276.

Mahadevan R, Bond DR, Butler JE *et al.* (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl Environ Microb* **72**: 1558–1568.

Médigue C & Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* **158**: 724–736.

Meyer P & Dworkin J (2007) Applications of fluorescence microscopy to single bacterial cells. *Res Microbiol* **158**: 187–194.

Nikolaev EV, Burgard AP & Maranas CD (2005) Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys J* **88**: 37–49.

Nogales J, Palsson BO & Thiele I (2008) A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Syst Biol* **2**: 79.

Notebaart RA, van Enckevort FHJ, Francke C, Siezen RJ & Teusink B (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**: 296.

Notebaart RA, Teusink B, Siezen RJ & Papp B (2008) Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol* **4**: e26.

Oberhardt MA, Puchałka J, Fryer KE, dos Santos VAPM & Papin JA (2008) Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *J Bacteriol* **190**: 2790–2803.

Oh Y-K, Palsson BO, Park SM, Schilling CH & Mahadevan R (2007) Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282**: 28791–28799.

Oliveira AP, Nielsen J & Förster J (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiology* **5**: 39.

Osterman A & Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol* **7**: 238–251.

Overbeek R, Begley T, Butler RM *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.

Pál C, Papp B & Lercher MJ (2005a) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**: 1372–1375.

Pál C, Papp B & Lercher MJ (2005b) Horizontal gene transfer depends on gene content of the host. *Bioinformatics* **21** (suppl 2): 222–ii223.

Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG & Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667–670.

Palsson BO (2006) *Systems Biology. Properties of Reconstructed Networks*. Cambridge University Press, New York, NY, USA.

Papin JA, Price ND, Edwards JS & Palsson BO (2002) The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J Theor Biol* **215**: 67–82.

Papin JA, Price ND, Wiback SJ, Fell DA & Palsson BO (2003) Metabolic pathways in the post-genome era. *Trends Biochem Sci* **28**: 250–258.

Papin JA, Stelling J, Price ND, Klamt S, Schuster S & Palsson BO (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol* **22**: 400–405.

Papp B, Pál C & Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**: 661–664.

Park JH, Lee SY, Kim TY & Kim HU (2008) Application of systems biology for bioprocess development. *Trends Biotechnol* **26**: 404–412.

Patil KR & Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *P Natl Acad Sci USA* **102**: 2685–2689.

Patil KR, Akesson M & Nielsen J (2004) Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotech* **15**: 64–69.

Patil KR, Rocha I, Förster J & Nielsen J (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics* **6**: 308.

Peng R-H, Xiong A-S, Xue Y *et al.* (2008) Microbial biodegradation of polyaromatic hydrocarbons. *FEMS Microbiol Rev* **32**: 927–955.

Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F & Schuster S (1999) METATOOL: for studying metabolic networks. *Bioinformatics* **15**: 251–257.

Pharkya P & Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng* **8**: 1–13.

Pharkya P, Burgard AP & Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* **14**: 2367–2376.

Poolman MG (2006) ScrumPy: metabolic modelling with Python. *Syst Biol (Stevenage)* **153**: 375–378.

Price ND, Reed JL & Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**: 886–897.

Raymond JW, Gardiner EJ & Willett P (2002) Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comput Sci* **42**: 305–316.

Reed JL & Palsson BO (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J Bacteriol* **185**: 2692–2699.

Reed JL & Palsson BO (2004) Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res* **14**: 1797–1805.

Reed JL, Vo TD, Schilling CH & Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4**: R54.

Reed JL, Famili I, Thiele I & Palsson BO (2006a) Towards multidimensional genome annotation. *Nat Rev Genet* **7**: 130–141.

Reed JL, Patel TR, Chen KH *et al.* (2006b) Systems approach to refining genome annotation. *P Natl Acad Sci USA* **103**: 17480–17484.

Ren Q, Kang KH & Paulsen IT (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32**: D284–D288.

Resendis-Antonio O, Reed JL, Encarnación S, Collado-Vides J & Palsson BØ (2007) Metabolic reconstruction and modeling of nitrogen fixation in *Rhizobium etli*. *PLoS Comput Biol* **3**: e192.

Risso C, Van Dien SJ, Orloff A, Lovley DR & Coppi MV (2008) Elucidation of an alternate isoleucine biosynthesis pathway in *Geobacter sulfurreducens*. *J Bacteriol* **190**: 2266–2274.

Rittmann BE (2008) Opportunities for renewable bioenergy using microorganisms. *Biotechnol Bioeng* **100**: 203–212.

Ro D-K, Paradise EM, Ouellet M *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**: 940–943.

Saghatelian A, Trauger SA, Want EJ, Hawkins EG, Siuzdak G & Cravatt BF (2004) Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* **43**: 14332–14339.

Saito N, Robert M, Kitamura S *et al.* (2006) Metabolomics approach for enzyme discovery. *J Proteome Res* **5**: 1979–1987.

Sauer U (2006) Metabolic networks in motion: ^{13}C-based flux analysis. *Mol Syst Biol* **2**: 62.

Schilling CH & Palsson BO (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J Theor Biol* **203**: 249–283.

Schilling CH, Edwards JS, Letscher D & Palsson BO (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng* **71**: 286–306.

Schilling CH, Covert MW, Famili I, Church GM, Edwards JS & Palsson BO (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J Bacteriol* **184**: 4582–4593.

Schneider G & Fechner U (2004) Advances in the prediction of protein targeting signals. *Proteomics* **4**: 1571–1580.

Schuetz R, Kuepfer L & Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* **3**: 119.

Schuster S, Fell DA & Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* **18**: 326–332.

Schuster S, Pfeiffer T & Fell DA (2008) Is maximization of molar yield in metabolic networks favoured by evolution? *J Theor Biol* **252**: 497–504.

Schwartz J-M, Gaugain C, Nacher JC, de Daruvar A & Kanehisa M (2007) Observing metabolic functions at the genome scale. *Genome Biol* **8**: R123.

Schwarz R, Liang C, Kaleta C *et al.* (2007) Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics* **8**: 313.

Segrè D, Vitkup D & Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *P Natl Acad Sci USA* **99**: 15112–15117.

Segrè D, Zucker J, Katz J *et al.* (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS* **7**: 301–316.

Senger RS & Papoutsakis ET (2008) Genome-scale model for *Clostridium acetobutylicum*: part I. Metabolic network resolution and analysis. *Biotechnol Bioeng* **101**: 1036–1052.

Serres MH, Goswami S & Riley M (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* **32**: D300–D302.

Shlomi T, Berkman O & Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *P Natl Acad Sci USA* **102**: 7695–7700.

Shlomi T, Eisenberg Y, Sharan R & Ruppin E (2007a) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* **3**: 101.

Shlomi T, Herrgard M, Portnoy V, Naim E, Palsson BØ, Sharan R & Ruppin E (2007b) Systematic condition-dependent annotation of metabolic genes. *Genome Res* **17**: 1626–1633.

Smallbone K, Simeonidis E, Broomhead DS & Kell DB (2007) Something from nothing – bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274**: 5576–5585.

Stelling J, Klamt S, Bettenbrock K, Schuster S & Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**: 190–193.

Stephanopoulos GN, Aristidou AA & Nielsen J (1998) *Metabolic Engineering. Principles and Methodologies*. Academic Press, Elsevier Science, San Diego, CA, USA.

Steuer R (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform* **7**: 151–158.

Stolyar S, Van Dien SJ, Hillesland KL, Pinel N, Lie TJ, Leigh JA & Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol* **3**: 92.

Stouthamer AH & Bettenhaussen C (1973) Utilization of energy for growth and maintenance in continuous and batch cultures of microorganisms. A reevaluation of the method for the determination of ATP production by measuring molar growth yields. *Biochim Biophys Acta* **301**: 53–70.

Sun J & Zeng A-P (2004) IdentiCS – identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics* **5**: 112.

Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien SJ, Keasling JD & Maranas CD (2007) Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. *Metab Eng* **9**: 387–405.

Suzuki N, Okai N, Nonaka H, Tsuge Y, Inui M & Yukawa H (2006) High-throughput transposon mutagenesis of *Corynebacterium glutamicum* and construction of a single-gene disruptant mutant library. *Appl Environ Microb* **72**: 3750–3755.

ter Kuile BH & Westerhoff HV (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett* **500**: 169–171.

Terzer M & Stelling J (2008) Large scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* **24**: 2229–2235.

Teusink B, van Enckevort FHJ, Francke C, Wiersma A, Wegkamp A, Smid EJ & Siezen RJ (2005) *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. *Appl Environ Microb* **71**: 7253–7262.

Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ & Smid EJ (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J Biol Chem* **281**: 40041–40048.

Thiele I, Vo TD, Price ND & Palsson BØ (2005) Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an *in silico* genome-scale characterization of single- and double-deletion mutants. *J Bacteriol* **187**: 5818–5830.

UniProt C (2008) The universal protein resource (UniProt). *Nucleic Acids Res* **36**: D190–D195.

Van Dien SJ & Lidstrom ME (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C(3) and C(4) metabolism. *Biotechnol Bioeng* **78**: 296–312.

van Winden WA, Heijnen JJ, Verheijen PJ & Grievink J (2001) A priori analysis of metabolic flux identifiability from (13)C-labeling data. *Biotechnol Bioeng* **74**: 505–516.

Varma A & Palsson BO (1994a) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microb* **60**: 3724–3731.

Varma A & Palsson BO (1994b) Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12**: 994–998.

von Kamp A & Schuster S (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* **22**: 1930–1931.

von Mering C, Jensen LJ, Kuhn M *et al.* (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35**: D358–D362.

Wiback SJ, Famili I, Greenberg HJ & Palsson BO (2004) Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J Theor Biol* **228**: 437–447.

Wiechert W (2001) ^{13}C metabolic flux analysis. *Metab Eng* **3**: 195–206.

Wiechert W, Möllney M, Isermann N, Wurzel M & de Graaf AA (1999) Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnol Bioeng* **66**: 69–85.

Wright J & Wagner A (2008) The systems biology research tool: evolvable open-source software. *BMC Syst Biol* **2**: 55.

Yang C, Hua Q & Shimizu K (2002) Integration of the information from gene expression and metabolic fluxes for the analysis of the regulatory mechanisms in *Synechocystis*. *Appl Microbiol Biot* **58**: 813–822.

Yang F, Qian H & Beard DA (2005) Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. *Metab Eng* **7**: 251–259.

Yang HH, Hu Y, Buetow KH & Lee MP (2004) A computational approach to measuring coherence of gene expression in pathways. *Genomics* **84**: 211–217.

Yeung M, Thiele I & Palsson B (2007) Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**: 363.

Yugi K, Nakayama Y, Kinoshita A & Tomita M (2005) Hybrid dynamic/static method for large-scale simulation of metabolism. *Theor Biol Med Model* **2**: 42.

Zamboni N, Kümmel A & Heinemann M (2008) anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9**: 199.

Zupke C & Stephanopoulos G (1994) Modeling of isotope distribution and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnol Prog* **10**: 489–498.

3.2.2 Compléments méthodologiques

L'article de revue ci-dessus est destiné à un lectorat peu familier des notions mathématiques. Nous compléterons donc notre présentation des modèles à base de contraintes dans les paragraphes qui suivent en explicitant le cadre mathématique sous-jacent aux hypothèses de modélisation présentées dans la revue.

Représentation des flux de réactions

L'état du métabolisme est décrit dans les modèles à base de contraintes uniquement par les flux de réactions. Pour un réseau donné de réactions, l'état du système est ainsi modélisé par un ensemble de nombre réels représentant chacun le flux d'une réaction du réseau. Cet ensemble de nombre est appelé *distribution de flux* et est manipulé mathématiquement sous la forme d'un vecteur (voir Figure 15).

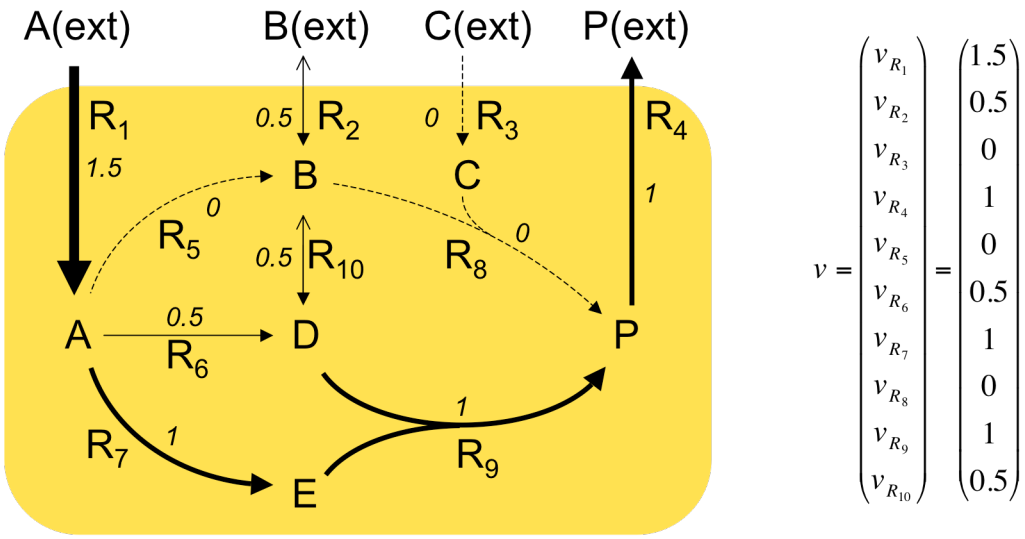


Figure 15. Représentation sous forme vectorielle d'une distribution de flux. À gauche, illustration d'un réseau métabolique théorique composé de 10 métabolites et 10 réactions. La zone jaune délimite le système (intérieur de la cellule par exemple). Les réactions R_1, R_2, R_3 et R_4 modélisent le transport des métabolites A, B, C et P entre l'extérieur et l'intérieur du système. Les valeurs des flux sont indiquées à proximité des réactions (en unités arbitraires) et illustrées par l'épaisseur de leurs flèches. À droite, représentation de la même distribution de flux sous forme vectorielle, telle qu'utilisée par les modèles à base de contraintes.

D'un point de vue physique, les flux de réactions manipulés représentent des valeurs moyennées sur des intervalles de temps s'étendant entre la seconde à la minute. L'échelle de temps choisie a une importance fondamentale pour ces modèles. Elle se situe en effet entre, d'une part, les temps de relaxation des cinétiques

enzymatiques (beaucoup plus rapides[34]) et, d'autre part, ceux des changements environnementaux considérés et des réponses régulatrices à ces changements (beaucoup plus lentes) (voir Figure 16). Cette échelle de temps permet donc d'étudier la répartition des flux métaboliques en réponse à différentes conditions environnementales ou de régulation tout en ignorant les dynamiques complexes d'ajustements rapides des cinétiques enzymatiques. Ces dernières peuvent être supposées être dans un état quasi-stationnaire, flux et concentrations métaboliques sont supposés constants (Stephanopoulos et al. 1998, pp.25-27, 82-83, 313-315). De plus, cette échelle de temps correspond relativement bien aux observations expérimentales typiquement réalisées : par exemple la mesure de taux de croissance ou de vitesse de consommation / production de métabolites externes.

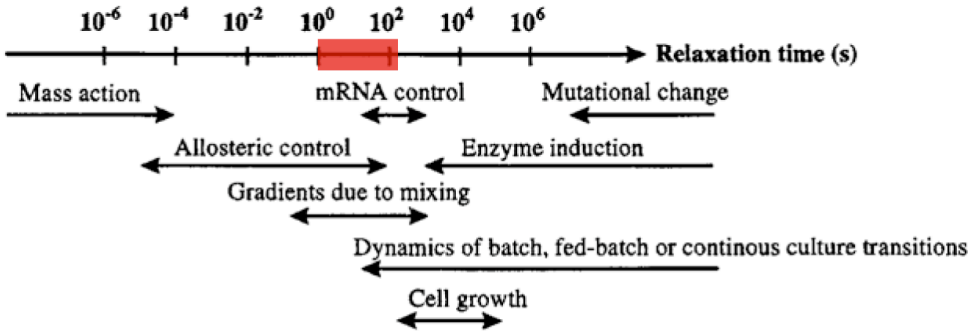


Figure 16. Temps de relaxations caractéristiques de différents processus cellulaires et du fonctionnement d'un bioréacteur. En rouge, temps caractéristiques considérés dans les modèles à base de contraintes. Adapté de (Stephanopoulos et al. 1998, p.25).

Lorsque aucune hypothèse n'est formulée sur le fonctionnement du métabolisme et qu'aucune contrainte n'est appliquée aux flux, ceux-ci peuvent prendre n'importe quelles valeurs réelles. L'ensemble des distributions de flux possibles dans le métabolisme est donc représenté par l'espace vectoriel $\mathbf{R}^n$ tout entier, où n est le nombre de réactions.

Expression mathématique des contraintes sur les flux

L'hypothèse d'état quasi-stationnaire se traduit dans le modèle par un ensemble de contraintes de *conservation de la masse* entre les flux.

[34] Notamment dans les conditions physiologiques, où les concentrations métaboliques sont faibles en regard des vitesses de réactions à cette échelle de temps (Stephanopoulos et al. 1998; Fell 1992).

Ces contraintes statuent que, pour chacun des métabolites dont les concentrations sont supposées stationnaires, les taux nets de formation sont nuls. En d'autres termes, les flux des réactions consommant ces métabolites compensent ceux des réactions les produisant. Cette contrainte n'est pas valide pour tous les métabolites. Certains subissent de manière significative la dilution due à la croissance des cellules, tandis que d'autres possèdent des concentrations non-stationnaires ou sont transformés par des processus non modélisés. L'ajout de réactions artificielles dans le modèle – réaction de biomasse dans le premier cas, réactions d'échange dans le deuxième (voir Figure 1 de l'article de revue) – permet de modéliser ces effets et d'appliquer correctement les contraintes de conservation de la masse à tous les métabolites.

L'utilisation d'une matrice stœchiométrique (voir Figure 1 de l'article de revue) permet d'appliquer les contraintes de conservation de la masse simultanément pour tous les métabolites. Le produit matriciel $S.v$, où S est la matrice stœchiométrique du réseau métabolique et v le vecteur de distribution de flux, calcule en effet directement le taux de formation de chacun des métabolites (voir Figure 17). Il en découle que l'équation matricielle $S.v=0$ définit directement l'ensemble des distributions de flux obéissant à la contrainte de conservation de la masse. Cette équation matricielle peut être vue comme un système d'équations linéaires entre les flux. Bien que celui-ci soit en général largement sous-déterminé, il contribue à préciser les distributions de flux possibles dans le réseau, en éliminant celles qui ne sont pas compatibles avec la conservation de la masse. En termes d'algèbre linéaire, l'application de cette contrainte réduit l'espace des distributions de flux de $\mathbf{R}^n$ à un de ses sous-espaces de dimension inférieure, $Ker(S)$, appelé le noyau de la matrice S.

$$
S.v = \begin{pmatrix}
1 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{pmatrix}
\cdot
\begin{pmatrix}
v_{R_1} \\ v_{R_2} \\ v_{R_3} \\ v_{R_4} \\ v_{R_5} \\ v_{R_6} \\ v_{R_7} \\ v_{R_8} \\ v_{R_9} \\ v_{R_{10}} \\ v_{E_{A(ext)}} \\ v_{E_{B(ext)}} \\ v_{E_{C(ext)}} \\ v_{E_{P(ext)}}
\end{pmatrix}
$$

$$
= \begin{pmatrix}
v_{R_1} - v_{R_5} - v_{R_6} - v_{R_7} \\
v_{R_2} + v_{R_5} - v_{R_8} - v_{R_{10}} \\
v_{R_3} - v_{R_8} \\
v_{R_6} - v_{R_9} + v_{R_{10}} \\
v_{R_7} - v_{R_9} \\
-v_{R_4} + v_{R_9} + v_{R_{10}} \\
-v_{R_1} - v_{E_{A(ext)}} \\
-v_{R_2} - v_{E_{B(ext)}} \\
-v_{R_3} - v_{E_{C(ext)}} \\
v_{R_4} - v_{E_{P(ext)}}
\end{pmatrix}
= \frac{d}{dt}
\begin{pmatrix}
c_A \\ c_B \\ c_C \\ c_D \\ c_E \\ c_P \\ c_{A(ext)} \\ c_{B(ext)} \\ c_{C(ext)} \\ c_{P(ext)}
\end{pmatrix}
= 0
$$

Figure 17. Matrice stœchiométrique et conservation de la masse. L'exemple reprend le réseau présenté Figure 15 augmenté de réactions d'échanges pour les métabolites extracellulaires (notés $E_{X(ext)}$ pour un métabolite X). Les concentrations métaboliques sont notées c_X.

Ces contraintes définissent principalement des relations entre les flux. Typiquement, deux réactions se suivant dans une voie métabolique sans branchement seront forcées par cette contrainte à se dérouler avec des flux égaux pour conserver la quantité du métabolite intermédiaire.

Toute information sur la valeur des flux est également utilisée pour contraindre le modèle. À ce sujet, nous avions montré dans l'article de revue qu'un large panel de types de connaissances pouvait se traduire directement par des contraintes sur les valeurs de flux. Il s'agit notamment de la réversibilité des réactions (flux uniquement positifs pour les réactions irréversibles), de leur inactivation (flux nul), d'hypothèse sur la valeur maximale de leurs flux (flux inférieurs à cette valeur), de valeurs mesurées (flux directement égal à cette valeur), et de la connaissance de l'environnement extérieur (flux des réactions d'échanges autorisés uniquement dans

le sens de l'excrétion pour les métabolites absents de l'environnement). Ces informations se traduisent dans le modèle simplement par des équations d'égalités ou d'inégalités sur les flux, $v_{min} \leq v \leq v_{max}$, qui viennent s'ajouter aux contraintes précédentes. Ces équations affinent encore plus l'ensemble des distributions de flux possibles dans le réseau. Notamment, les contraintes de conservation de la masse « propagent » dans le réseau des connaissances locales de valeurs de flux. D'un point de vue ensembliste, l'ajout de ces inégalités réduit le noyau de la matrice stœchiométrique à un ensemble qui n'est plus un espace vectoriel, mais possède les propriétés de linéarité et de convexité. Les ensembles ayant ces propriétés ont été largement étudiés et un grand panel d'outils permettant de les explorer a été développé et regroupé sous l'appellation *analyse convexe* (Rockafellar 1970).

Les contraintes engendrant des ensembles convexes de distribution de flux sont majoritairement utilisées dans les modèles métaboliques à base de contraintes, probablement grâce à la simplicité de leur utilisation. Elles suffisent à décrire les hypothèses fondamentales de modélisation permettant de prédire les phénotypes de croissance. Dans nos travaux, nous nous sommes donc limités à ces types de contraintes. Comme évoqué dans l'article de revue, toute hypothèse ou connaissance pouvant être traduite en contrainte sur les flux peut en principe être intégrée au modèle. Des travaux ont été effectués dans ce sens pour tenir compte de la régulation (Covert et al. 2001), de la signalisation (Lee et al. 2008), des lois de la thermodynamique (Beard et al. 2002; Beard et al. 2004) ou de mesures de concentrations métaboliques (Kümmel et al. 2006a; Henry et al. 2007) ; ils résultent néanmoins en des contraintes mathématiques non linéaires et introduisant parfois des variables entières. Toute la difficulté réside alors dans l'exploitation mathématique de ces contraintes et la recherche des distributions de flux compatibles avec elles.

Exploration des états métaboliques

Les outils développés pour l'analyse convexe linéaire s'appliquent directement aux modèles à base de contraintes lorsque les contraintes s'y prêtent. Nous en avons évoqué trois classes dans l'article de revue : l'exploration de l'ensemble des distributions de flux par échantillonnage (voir à ce propos une revue encore plus récente par Schellenberger & Palsson (2009)), description de modes élémentaires et optimisation. Nous ne reviendrons pas sur les deux premières méthodes ; elles sont

particulièrement bien adaptées à la caractérisation de l'ensemble des distributions de flux, mais sont coûteuses en temps de calcul. Les méthodes d'optimisations, bien que ne recherchant que des distributions de flux particulières, sont bien plus performantes et également utilisables pour prédire des phénotypes de croissance.

L'essor de la recherche opérationnelle a stimulé le développement de méthodes d'optimisation sous contraintes efficaces (Boyd & Vandenberghe 2004). Ces méthodes ont typiquement pour but de résoudre des problèmes du type :

$$\begin{aligned} &\textit{maximiser } f_0(x), \textit{ tel que :} \\ &\quad f_i(x) \leq b_i \textit{ pour } i = 1,\ldots,m \end{aligned}$$

où le vecteur $x = (x_1,\ldots,x_n)$ est la variable à optimiser, $f_0 : \mathbf{R}^n \rightarrow \mathbf{R}$ la fonction objectif, les $f_i : \mathbf{R}^n \rightarrow \mathbf{R}$ des fonctions de contraintes, et les b_i les bornes des contraintes. Lorsque les fonctions de contraintes et d'objectif sont linéaires, ces problèmes sont dits de *programmation linéaire*. Des méthodes et des logiciels particulièrement performants existent pour les résoudre[35] (Boyd & Vandenberghe 2004; Moisdon 2000; Löfberg 2004).

La linéarité des contraintes sur les valeurs de flux et de conservation de la masse permet d'exploiter ces outils. Leur application la plus courante est la prédiction de phénotype de croissance par la méthode de « Flux Balance Analysis » (FBA) qui consiste à maximiser le flux de la réaction de biomasse dans le modèle (c'est une fonction objectif évidemment linéaire, le flux de la réaction de biomasse étant directement une variable du problème) :

$$\begin{aligned} &\textit{maximiser } v_{Rbiomasse}\textit{ , tel que :} \\ &\quad S.v = 0 \\ &\quad v_{min} \leq v \leq v_{max} \end{aligned}$$

où $v_{Rbiomasse}$ est la variable correspondant au flux de la réaction de biomasse.

La résolution de ce problème explore les distributions de flux compatibles avec les contraintes et en fournit une permettant d'atteindre un flux maximal pour la réaction

[35] Voir par exemple un inventaire sur le Wiki des serveurs d'optimisation NEOS : http://wiki.mcs.anl.gov/NEOS/

de biomasse. Elle permet donc d'évaluer les capacités du réseau métabolique en terme de production de biomasse.

Bien que ne caractérisant pas l'ensemble des distributions de flux possibles, les méthodes d'optimisation permettent néanmoins d'interroger le modèle métabolique au coup par coup quant à ses capacités à remplir un objectif donné. Dans le cas présenté ci-dessus, il s'agit de l'aptitude à croître (c.-à-d. à former de la biomasse), mais, utilisée astucieusement, l'optimisation permet d'explorer de nombreuses autres caractéristiques du réseau métabolique (Price et al. 2004). Nous mentionnerons par exemple l'étude de contraintes de couplages métaboliques entre réactions (Burgard et al. 2004), le calcul de plages de flux possibles pour chaque réaction (Mahadevan & Schilling 2003) ou, à l'inverse, la recherche d'un objectif métabolique représentant au mieux des distributions de flux observées (Burgard & Maranas 2003; Schuetz et al. 2007).

3.3 Modélisation du métabolisme et phénotypes de croissance: état de l'art

Dans cette dernière partie introductive sur la modélisation du métabolisme, nous effectuerons un état de l'art – à la date du début de la thèse, fin 2005 – des travaux permettant d'interpréter les phénotypes de croissance à l'aide de modèles du métabolisme.

3.3.1 Modèles à base de graphe

Malgré leur aisance à représenter un réseau métabolique cellulaire dans son ensemble, les graphes métaboliques ont été relativement peu utilisés pour relier phénotypes de croissance et métabolisme. Ces derniers se sont en effet révélés bien adaptés à étudier l'organisation et la structure des réseaux métaboliques, mais beaucoup moins à aborder leur fonctionnement en lui-même. De nombreux travaux cherchant à interpréter des données d'essentialités à la lumière de graphes de réseaux biologiques[36] se sont en réalité limités à établir des corrélations entre caractéristiques topologiques (ex. la « centralité » d'un gène) et importance phénotypique (Jeong et al. 2001; Batada et al. 2006; Hahn & Kern 2005).

[36] Principalement pour les réseaux d'interactions protéines-protéines d'ailleurs.

Quelques initiatives ont exploité les méthodes d'expansion de réseau pour prédire l'effet du changement d'environnement ou de la délétion de gènes sur la viabilité des cellules (Handorf et al. 2005; Wunderlich & Mirny 2006). Pour ce faire, ces travaux examinèrent si les métabolites nécessaires à la survie de la cellule pouvaient être tous synthétisés par le graphe métabolique perturbé par les délétions, à partir des métabolites de l'environnement. Ces initiatives, bien qu'obtenant des performances de prédictions relativement proches de celles des modèles à base de contraintes, restèrent relativement isolées. Il est probable que l'aspect quantitatif des modèles à base de contraintes et leur égale facilité de mise en œuvre ont favorisé leur utilisation.

3.3.2 Modèles à base de contraintes

Les modèles à base de contraintes furent très rapidement utilisés pour prédire des phénotypes de croissance, dès lors que des réseaux d'échelles cellulaires furent reconstruits. Ce type d'application contribua d'ailleurs fortement à populariser ce cadre de modélisation[37] (Edwards & Palsson 2000; Edwards et al. 2001). Nous avons déjà longuement évoqué les travaux relatifs à la prédiction de phénotypes de croissance dans l'article de revue, aussi nous contenterons-nous ici de rappeler ceux ayant été effectués avant 2006 et en rapport avec notre thèse. Ils se répartissent en deux types de contributions : celles d'ordre méthodologique et celles confrontant les prédictions aux phénotypes expérimentaux.

Les méthodes de prédiction des phénotypes de croissance furent déjà largement explorées fin 2005. À l'utilisation « classique » de la méthode FBA pour prédire les phénotypes de mutants de délétions (Varma & Palsson 1994) vinrent s'ajouter les méthodes MoMA (Segrè et al. 2002) et ROOM (Shlomi et al. 2005) ayant pour but de corriger l'hypothèse de fonctionnement optimal du métabolisme. Dans le même ordre d'idée, Imielinski et al (2005) proposa une méthode basée sur la productibilité des métabolites pour associer la létalité des délétions à des métabolites, et définir un ensemble de métabolites essentiels. Enfin, Klamt & Gilles (2004) proposèrent pour la première fois une méthode permettant d'énumérer de manière exhaustive les délétions, simples ou multiples, létales pour l'organisme.

[37] Signe de l'intérêt grandissant pour ce type de modélisation, plusieurs groupes avaient publié, notamment en 2005, des reconstructions globales de modèles métaboliques (voir Table 4 de l'article de revue).

La fiabilité des phénotypes de croissance prédits par les modèles à base de contraintes fut démontrée par un grand nombre de travaux confrontant prédictions et observations expérimentales. Cette démarche fut effectuée pour évaluer à la fois la validité des prédictions quantitatives de taux de croissance (Edwards et al. 2001; Ibarra et al. 2002; Duarte, Palsson et al. 2004) et celle des prédictions des phénotypes qualitatifs de croissance pour des mutants de délétion (Edwards & Palsson 2000; Famili et al. 2003; Covert et al. 2004; Duarte, Herrgard et al. 2004; Borodina et al. 2005; Kuepfer et al. 2005; Thiele et al. 2005). Dans la majorité de ces travaux, les observations expérimentales furent simplement utilisées pour évaluer les performances de prédiction des modèles. Néanmoins, les travaux les plus récents évoquèrent l'idée d'exploiter les observations expérimentales faussement prédites pour améliorer la connaissance du métabolisme : Duarte, Herrgard et al. (2004) ont, dans le cas de la levure, classé les fausses prédictions par type de cause probable, et Covert et al. (2004) ont corrigé des interactions régulatrices dans un modèle mixte métabolisme-régulation d'*E. coli* sur la base de données expérimentales de phénotypes de croissance et d'expression transcriptionnelle.

4 Notre organisme modèle : *Acinetobacter baylyi* ADP1

4.1 Caractéristiques remarquables

Tout au long des travaux de cette thèse, nous avons utilisé la bactérie *Acinetobacter baylyi* ADP1 comme organisme modèle. Le choix de cet organisme fut guidé par la réalisation au sein même du Genoscope d'un vaste projet expérimental d'exploration de son métabolisme. Ce projet fut initié en 2002 par le séquençage et l'annotation de son génome et poursuivi par la création et l'étude d'une collection de mutants de délétion pour chacun de ses gènes (projet nommé « Thesaurus métabolique » (de Berardinis et al. 2008)). La taille de la collection de mutants et la capacité de tester leurs phénotypes de croissance sur divers environnements ont constitué une ressource expérimentale précieuse à l'application des méthodes de modélisation développées dans cette thèse. De plus, les reconstructions du réseau et du modèle métabolique d'*A. baylyi* ont participé à l'interprétation des résultats expérimentaux en apportant des outils d'analyse et de visualisation des processus métaboliques identifiés chez la bactérie.

A. baylyi ADP1 est une γ-protéobactérie Gram négative, oxidase négative et strictement aérobie du genre *Acinetobacter*, lui même appartenant à l'ordre des *Pseudomonales* (voir Figure 18 B). La classification précise de cette souche – précédemment désignée par BD413 et parfois anciennement classée dans l'espèce *Acinetobacter calcoaceticus* – est très récente (Vaneechoutte et al. 2006). La caractérisation de l'espèce *A. baylyi* remonte d'ailleurs seulement à 2003 (Carr et al. 2003). Historiquement, cette souche est issue des travaux de Taylor et Juni (1961a; 1961b; 1961c) portant sur la synthèse des capsules de polysaccharides chez les procaryotes. Ceux-ci avaient isolé par enrichissement sur un milieu contenant du butane-2,3-diol comme seule source de carbone une bactérie du sol produisant de grandes quantités d'exopolysaccharides, qu'ils désignèrent par *Acinetobacter calcoaceticus* BD4 (Taylor & Juni 1961a). Une étude ultérieure par mutagenèse conduisit Juni et Janik (1969) à mettre en évidence l'aptitude naturelle à la transformation de cette souche, et à en dériver la souche BD413 produisant une capsule polysaccharidique très réduite facilitant sa manipulation. *A. baylyi* ADP1 provient directement de cette dernière souche. Dans la suite de ce manuscrit, nous désignerons simplement par *A. baylyi* la souche ADP1 lorsqu'il n'y aura pas d'ambiguïté.

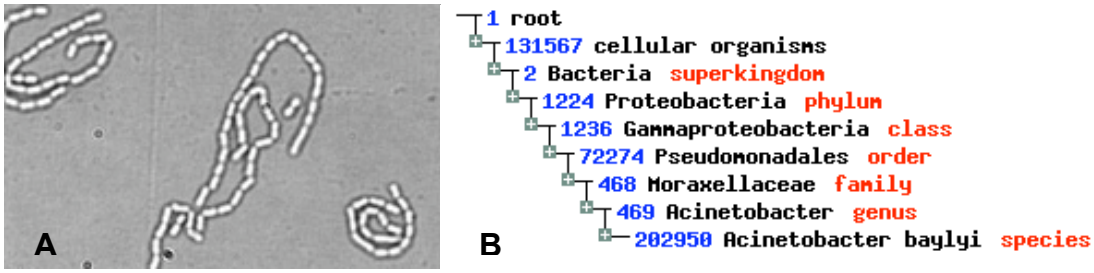


Figure 18. *Acinetobacter baylyi.* **A** Vue au microscope d'une population d'*Acinetobacter baylyi*. **B** Classification taxonomique d'*A. baylyi* . En bleu sont indiqués les identifiants taxonomiques du NCBI (extrait de http://srs.ebi.ac.uk).

La caractéristique la plus remarquable d'*A. baylyi* est son exceptionnelle aptitude naturelle à la transformation par de l'ADN exogène (aussi bien circulaire que linéaire) : cette souche est naturellement compétente[38] (Juni 1972; Palmen & Hellingwerf 1997) et effectue facilement des recombinaisons homologues (de Vries & Wackernagel 2002). Cette aptitude est *a priori* spécifique aux souches de l'espèce

[38] La compétence désigne la capacité d'une cellule à importer de l'ADN extracellulaire.

baylyi ; il a été montré que d'autres espèces d'*Acinetobacter* effectuent des transformations naturelles, mais à des fréquences 100 à 1000 fois moindre (Vaneechoutte et al. 2006). Alors que la question de l'avantage sélectif procuré par cette aptitude dans un environnement naturel est encore largement ouverte (Young et al. 2005, p.522), elle offre aux expérimentateurs un excellent outil de travail génétique, pouvant avantageusement remplacer le modèle habituel *Escherichia coli* K12[39]. Les possibilités expérimentales offertes sont nombreuses, en particulier : les délétions simples ou multiples de séquences chromosomiques (notamment de gènes), l'expression de gènes hétérologues (promouvant *A. baylyi* comme support à la construction de nouveaux systèmes métaboliques), l'étiquetage de séquences génétiques (« gene tagging »), ou le remplacement de gènes par des variants mutés (Metzgar et al. 2004; Young et al. 2005).

Une autre caractéristique remarquable d'*A. baylyi*, partagée cette fois-ci avec d'autres membres de l'ordre des *Pseudomonales*, est sa grande polyvalence nutritionnelle. Étant une bactérie du sol, celle-ci est capable d'utiliser comme seules sources de carbone et d'énergie un large panel de composés présents dans ce type d'environnement, notamment des molécules aromatiques et des acides organiques à longues chaînes (Young et al. 2005; Barbe et al. 2004) (voir plus loin Tableau 4). Ces types de molécules sont en effet couramment produits par le métabolisme des plantes. Les voies métaboliques de dégradation de ces composés chez *A. baylyi* ont été et font d'ailleurs toujours l'objet de recherches relativement nombreuses, que ce soit sur les conversions biochimiques elles-mêmes (Williams & Ray 2008) ou leurs régulations (Gerischer et al. 2008). À l'inverse, et notamment par comparaison à *Escherichia coli* dont l'environnement naturel se situe dans l'intestin d'animaux à sang chaud (Neidhardt 1996), peu de sucres sont métabolisés par les bactéries du genre *Acinetobacter*.

[39] *A. baylyi* ADP1 est en effet 10 à 100 fois plus compétente que des cellules *Escherichia coli* rendues compétentes par traitement au chloride de calcium (Metzgar et al. 2004, p.5781).

La souche *A. baylyi* ADP1 est non pathogène[40] ; une caractéristique qui n'est pas commune à toutes les espèces du genre *Acinetobacter*. En effet, certaines d'entre elles – notamment *Acinetobacter baumanii* – sont impliquées dans des cas d'infections nosocomiales[41] et opportunistes, parfois de manière épidémique (Abbott 2005; Bergogne-Bérézin & Towner 1996). Ces souches ont en outre développé des résistances à de multiples antibiotiques, rendant le traitement des infections particulièrement problématique (Fournier et al. 2006). La non-pathogénicité d'*A. baylyi* la rend manipulable en laboratoire sans contrainte particulière et sa proximité phylogénétique aux espèces pathogènes peut en faire un modèle d'étude pour ces dernières sur certains aspects, comme cela a été le cas pour l'analyse de leurs génomes (Fournier et al. 2006; Vallenet et al. 2008).

En dehors des études académiques, les bactéries du genre *Acinetobacter* et *A. baylyi* en particulier occupent une place non négligeable dans les applications biotechnologiques. Celles-ci sont en effet déjà utilisées pour dégrader des polluants, tels que des biphényls, phénols, benzoates, nitriles ou du pétrole brut, et produire des composés biochimiques, par exemple des lipases, protéases, bioémulsifiants, de cyanophycine et différents types de biopolymères (Gutnick & Bach 2008; Abd-El-Haleem 2003). La manipulation génétique aisée d'*A. baylyi* et l'intérêt croissant porté à son métabolisme sont propices à favoriser de nouvelles utilisations industrielles dans le futur.

4.2 Annotation du génome

Le Genoscope a entrepris en 2002 de séquencer et de réaliser une annotation détaillée du génome d'*A. baylyi*. Ce travail fut achevé et publié en 2004 (Barbe et al. 2004). Un effort relativement conséquent fut consacré à l'annotation ; faisant suite à une étape de prédiction automatique des gènes et de leurs fonctions, l'annotation de chacun d'entre eux fut complétée et validée manuellement au regard des

[40] Une étude récente de Chen *et al* (2008) attribue des cas d'infections nosocomiales à des souches de l'espèce *A. baylyi*, à partir d'analyses de leurs ARNr 16S. Le séquençage ultérieur de ces souches a cependant montré des dissimilarités notables avec *A. baylyi* ADP1 (B. Chen et V. de Berardinis, communication personnelle).
[41] Infections dont la source se situe en milieu hospitalier.

connaissances antérieures sur la biologie *d'A. baylyi* et de leurs contextes génomiques (Barbe et al. 2004; Vallenet et al. 2006).

Le génome d'*A. baylyi* compte 3,6 millions de paires de bases et affiche une composition en bases G et C de 40,4%. Ces caractéristiques le distinguent notablement de ceux des bactéries pourtant proches *Pseudomonas putida* KT2440 et *Pseudomonas aeruginosa* PAO1 (tailles d'environ 6,3 Mpb et composition en GC de 62-67%). Début 2009, son annotation comptait 3309 séquences codantes, incluant 3206 gènes validés et 103 annotés comme probablement non fonctionnels (comprenant des pseudo-gènes, des gènes à séquences très courtes ou codant à faible probabilité). La Figure 19 présente certaines des caractéristiques du génome sur une vue circulaire.

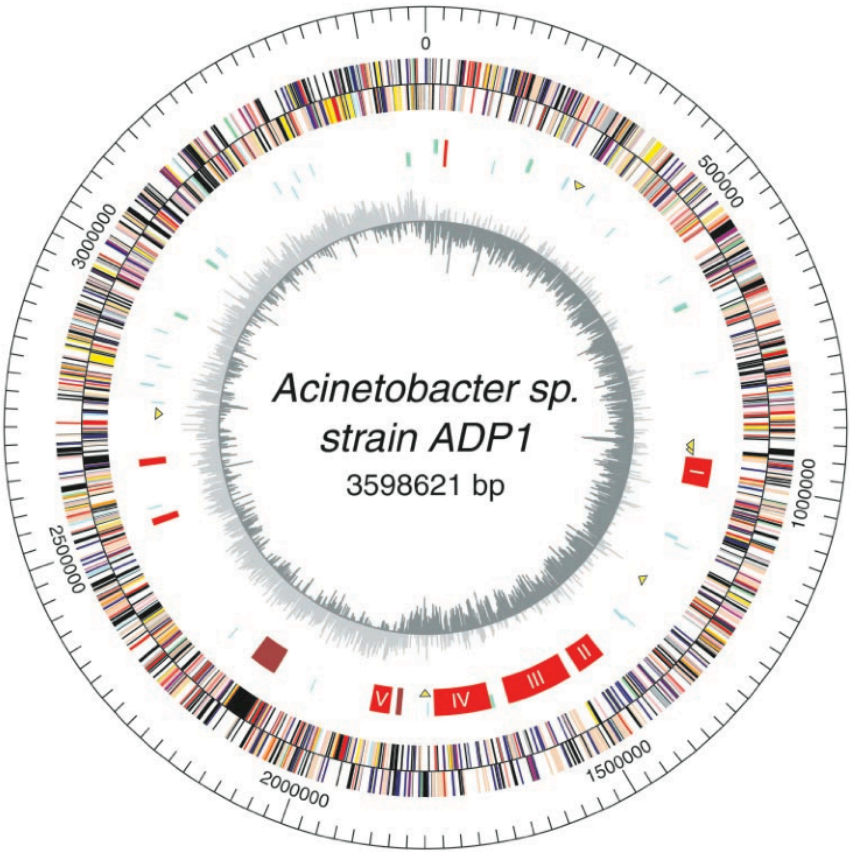


Figure 19. **Vue circulaire du génome d'*A. baylyi*.** Légende des cercles, de l'intérieur vers l'extérieur : (1) biais GC (G+C)/(G-C), (2) îlots cataboliques (rouge), région phagique (marron), éléments transposables (triangles jaunes), ARNt (bleu), ARNr (vert), et (3) gènes prédits dans les sens antihoraire (intérieur) et horaire (extérieur) colorés leur appartenance à des catégories de fonctions : saumon, biosynthèse des acides aminés ; bleu clair, biosynthèse des cofacteurs ; vert clair, enveloppe cellulaire ; rouge, processus cellulaires; jaune, métabolisme de l'ADN ; vert, métabolisme énergétique ; violet, métabolisme des acides gras et phospholipides ; rose, synthèse des protéines ; orange, synthèse des nucléotides ; bleu, fonctions de régulation ; gris, transcription ; turquoise, protéines de transport et d'interaction ; noir, protéines hypothétiques. Extrait de Barbe *et al* (2004).

L'analyse du génome d'*A. baylyi* corrobora les caractéristiques de son métabolisme connues précédemment (Barbe et al. 2004). Peu de gènes associés à des voies de dégradation de sucres furent découverts dans le génome. Cependant, le nombre de gènes impliqués dans le catabolisme de composés chimiques divers – principalement des aromatiques et de longs acides organiques – est remarquablement élevé (voir Tableau 4). L'examen de l'organisation de ces gènes sur le chromosome a confirmé par ailleurs qu'une grande partie d'entre eux se regroupent dans des régions chromosomiques précises, appelées îlots cataboliques (Doten et al. 1987) (voir Tableau 4 et Figure 19). Le but de ces regroupements n'est pas élucidé ; une hypothèse émise à ce sujet est la suivante : leur localisation dans le deuxième quadrant leur permettrait de bénéficier des duplications fréquentes du chromosome à cet endroit. La présence en plusieurs copies des gènes serait alors favorable à leur expression accrue, leur évolution (une copie évolue tandis que l'autre assure le maintien de la fonction initiale), voire leur échange avec d'autres organismes par transformation naturelle (Reams & Neidle 2004). En dehors des processus cataboliques, l'annotation du génome d'*A. baylyi* a permis d'élucider une part significative des voies de biosynthèses et de génération d'énergie, comme nous le verrons dans le chapitre consacré à la reconstruction de son réseau métabolique.

Compounds	Final products	Gene name	ACIAD number	Cluster
Alkanesulfonates	Sulfite + aldehyde	ssu	0034–0038	—
Vanillate	Protocatechuate	van[a]	0978–0983; 0988	I
Betaine	Betaine aldehyde	bet	1008–1012	I
Acetoin	Acetate	aco	1014; 1017–1022	I
Urea	Ammonia	ure	1088–1096	—
Salicylate esters	Catechol	sal[a]	1424–1427	II
Aryl esters	Catechol	are[a]	1428–1431	II
Benzoate	Catechol	ben[a]	1433–1440	II
Catechol	Succinate + acetyl CoA	cat[a]	1442–1451	II
Alkanesulfonates	Sulfite + aldehyde	ssu-like	1505; 1518; 1527; 1535	III
Dibenzothiophene	Hydroxybiphenyl + sulfite	sox	1510–1512	III
Sulfuric esters	Sulfate + phenol/alcohol	ats	1586–1588; 1592–1593; 1600–1601	III
Nitriles	Aliphatic amides	nth	1614–1616; 1622	III
Amidase	Ammoniac + acid	amdA	1618	III
Dicarboxylic acids	Succinate + acetyl CoA	dca[a]	1684; 1688–1698	IV
Protocatechuate	Succinate + acetyl CoA	pca[a]	1702–1712	IV
Quinate	Protocatechuate	qui[a]	1713–1716	IV
p-Hydroxybenzoate	Protocatechuate	pob[a]	1717–1719	IV
Chlorogenate	Quinate + caffeate	hca[a]	1720	IV
Caffeate	Protocatechuate	hca[a]	1722–1728	IV
Malonate	Acetate	mdc	1753–1762	IV
Nitrate/nitrite	Ammonia	nas	1908–1914	V
Sarcosine	Glycine + formaldehyde	sox	2550–2552	—
Anthranilate	Catechol	ant[a]	2669–2671	—
Alkanesulfonate	Sulfite + aldehyde	msu	3470–3472	—

[a]Genes already identified in *Acinetobacter* sp. ADP1.

Tableau 4. Composés pouvant être catabolisés par *A. baylyi* et gènes associés aux processus de dégradation. ACIAD désigne les identifiants uniques des gènes d'*A. baylyi*. Extrait de Barbe *et al* (2004)

Le déchiffrage de la séquence génomique d'*A. baylyi* permit d'obtenir une vision globale des activités biochimiques participant à son métabolisme. De plus, la

séquence ainsi que les annotations attenantes constituent une ressource de valeur pour les projets de recherche sur cet organisme et ceux proches. En attestent les 67 travaux référençant à ce jour l'annotation du génome[42]. L'interface MAGE met l'annotation de ce génome à la disposition de tout utilisateur et la maintient à jour par rapport aux gènes nouvellement annotés (Vallenet et al. 2006).

4.3 Collection de mutants de délétion

La simplicité des manipulations génétiques d'*A. baylyi* offerte par sa compétence naturelle en fait un organisme idéal pour l'application à haut débit de méthodes d'investigation génétique (Metzgar et al. 2004). Pour cette raison, et dans le but notamment d'identifier de nouvelles fonctions enzymatiques, l'équipe Thesaurus du Genoscope a construit une collection de mutants de délétion pour chacun des gènes d'*A. baylyi* (de Berardinis et al. 2008). Ce travail s'inscrit directement dans la veine des travaux de génétique à haut débit présentés ci-dessus (voir partie 2.2).

La technique de délétion utilisée tire naturellement parti des capacités de transformation et de recombinaison d'*A baylyi*. Elle consiste à remplacer via une recombinaison homologue le gène ciblé par une cassette d'intégration contenant un gène de résistance à un antibiotique (la kanamycine), permettant de sélectionner sur l'antibiotique les clones ayant recombiné correctement[43] (voir Figure 20). L'insertion correcte dans le locus ciblé est ensuite validée par une série de PCR qui amplifient des fragments d'ADN situés à cheval entre le génome et le gène de résistance (PCR entre amorces P1 et P6, P3 et P2, et P7 et P8, voir Figure 20). L'obtention de fragments amplifiés de longueurs attendues confirme alors indirectement l'intégration du gène de résistance au bon endroit dans le génome.

[42] Nombre d'articles citant l'article de Barbe *et al* (2004) au 27 février 2009 d'après ISI Web of Knowledge (Thomson Reuters, Inc.).
[43] La méthode de création de la cassette d'intégration est basée sur la technique de « spliced PCR » (Murphy et al. 2000) adaptée à *A. baylyi* par Metzgar et al (2004).

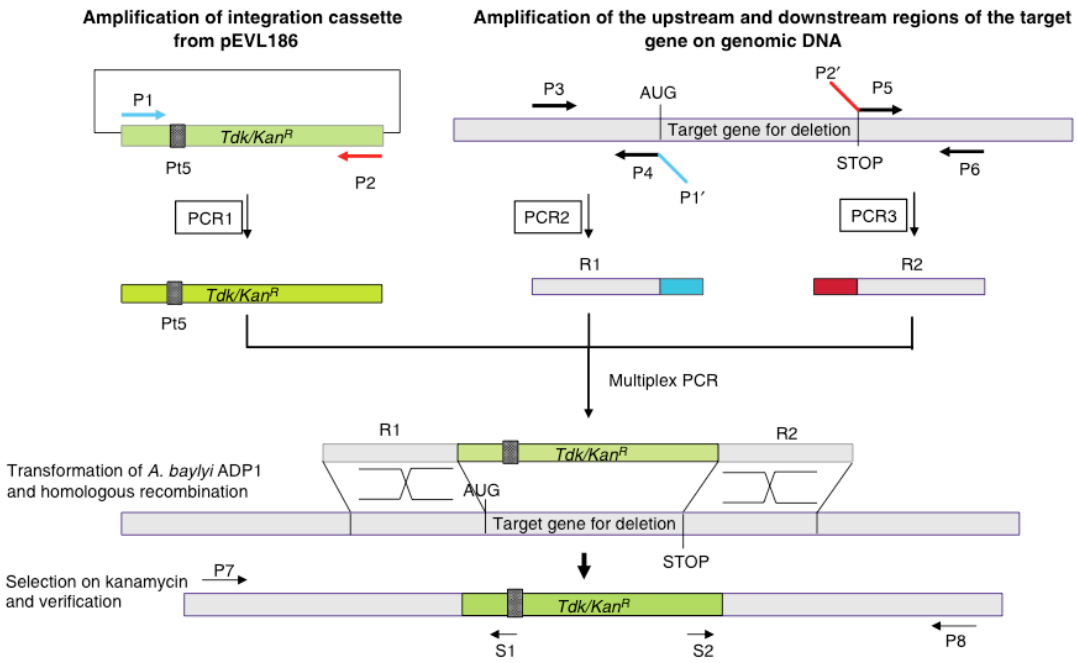


Figure 20. Technique de remplacement d'un gène d'*A. baylyi* par recombinaison homologue. Tout d'abord, une cassette de remplacement est créée en encadrant un gène de résistance à la kanamycine (Kan^R) muni d'un promoteur fort (Pt5) par des séquences homologues aux extrémités du gène à remplacer (R1 et R2). L'assemblage de la cassette est réalisé par une succession de PCR. *A. baylyi* est ensuite transformée avec cette cassette et mise en culture sur un milieu minimal (avec du succinate comme seule source de carbone) contenant de la kanamycine, permettant de sélectionner les clones ayant intégrés la cassette dans leur génome. Diverses PCR sont finalement réalisées sur les clones sélectionnés pour vérifier la bonne intégration de la cassette en remplacement du gène ciblé (vérification des longueurs des fragments entre les amorces P7 et P8, P1 et P6, et p3 et p2). Figure extraite de de Berardinis *et al* (2008)

Cette étape de vérification a mis en évidence, au cours des expériences de création des mutants, un phénomène de duplication de grandes régions chromosomiques perturbant la délétion du gène ciblé. En effet, pour les clones de certains gènes ciblés, les PCR d'amorces P7 et P8 (voir Figure 20) amplifièrent des fragments d'ADN de deux longueurs différentes, l'une correspondant à la région génomique possédant le gène ciblé intact et l'autre à la région génomique ayant intégré le gène de résistance (clones appelés « doubles bandes »). Une approche par puce CGH[44] montra alors que, dans le cas de ces clones, de larges régions chromosomiques incluant les gènes ciblés avaient effectivement été dupliquées, autorisant conjointement l'intégration du gène

[44] CGH : « Comprehensive Genomic Hybridization array ». Puce d'hybridation génomique comparative permettant de comparer, pour deux sources d'ADN distinctes, le nombre de copies des fragments d'ADN correspondant aux oligonucléotides (*sondes*) placés sur la puce. Dans le cas d'*A. baylyi,* des sondes couvrant l'ensemble du génome furent placées sur les puces et les quantités d'ADN furent comparées entre une souche sauvage et le clone mutant testé.

de résistance et la conservation d'une copie du gène cible (de Berardinis et al. 2008). Toutefois, malgré l'échec apparent de la manipulation génétique, ces gènes ciblés peuvent être considérés comme contribuant de manière significative à la survie de la bactérie, voire comme étant essentiels. En effet, la fréquence d'occurrence de telles duplications est rare (Reams & Neidle 2004) ; il est donc peu probable que des clones présentant ces duplications aient été retenus lors de l'expérience sans que ces duplications, et donc la conservation du gène ciblé, ne leur confère un avantage sélectif significatif. L'interprétation d'essentialité de ces gènes a de plus été corroborée par le fait qu'une grande majorité de leurs gènes homologues chez *Escherichia coli* et *Pseudomonas aeruginosa* sont également essentiels (de Berardinis et al. 2008).

Le processus de délétion fut appliqué à l'ensemble quasiment complet des gènes d'*A. baylyi*. Parmi eux 2594 donnèrent lieu à des mutants viables, correspondant à des gènes non-essentiels, et 499 à des mutants non viables ou « doubles bandes », correspondant aux gènes considérés comme essentiels. La totalité des 2594 mutants viables a été conservée et constitue la collection de mutants d'*A. baylyi*.

L'examen des catégories fonctionnelles liées aux gènes essentiels montre que ceux-ci composent une partie significative des voies métaboliques de biosynthèse (voir Figure 21). Ce résultat est en accord avec la composition du milieu choisi pour sélectionner les mutants. En effet, ce milieu minimal ne disposant que du succinate comme source de carbone, le bon fonctionnement des voies de biosynthèses est essentiel pour permettre aux mutants de disposer des métabolites nécessaires à leur survie.

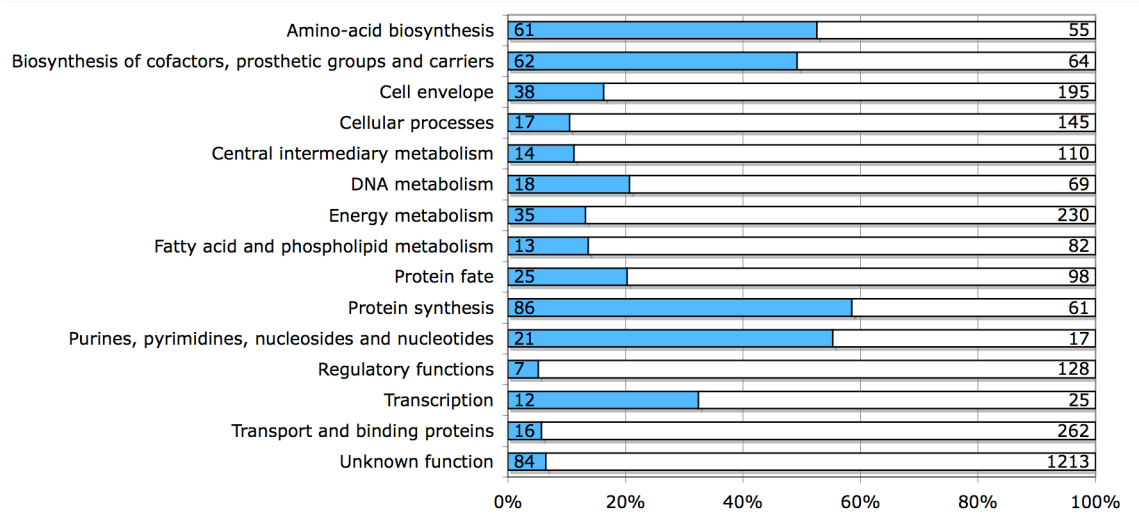


Figure 21. Proportion de gènes essentiels pour chaque catégorie fonctionnelle de la classification *TIGR Roles*. En bleu, gènes essentiels ; en blanc, gènes non-essentiels. Les catégories ne sont pas exclusives, certains gènes appartiennent à plusieurs d'entre elles.

Plusieurs facteurs de différences ont été mis en évidence par la comparaison des essentialités avec les gènes homologues chez *Pseudomonas aeruginosa* et *Escherichia coli* (de Berardinis et al. 2008). Dans le cas de *P. aeruginosa*, le principal facteur de différence est le milieu de sélection des mutants ; celui-ci est un milieu riche (LB) contenant de nombreux métabolites rendant non nécessaires des voies de biosynthèse (Liberati et al. 2006). De plus, la comparaison a révélé des gènes faussement identifiés comme essentiels chez *P. aeruginosa* du fait de l'utilisation d'une technique de transposon, moins fiable que la délétion ciblée (voir partie 2.2). Ces deux facteurs ne sont pas présents dans la comparaison aux essentialités d'*E. coli* ; la technique utilisée est similaire à celle d'*A. baylyi* et les mutants ont été phénotypés sur un milieu minimal (glucose) (Baba et al. 2006). Les différences d'essentialité entre *A. baylyi* et *E. coli* sont le reflet de différences dans leur métabolisme énergétique (la capacité anaérobie d'*E. coli* rend non-essentielles certaines sous-unités de l'ATP synthase), de la présence d'isoenzymes ou de voies métaboliques alternatives dans une seule des deux bactéries, ou de réelles différences dans certaines voies métaboliques (de Berardinis et al. 2008). En complément de ces analyses, la confrontation des voies métaboliques connues pour *A. baylyi* aux essentialités de gènes permit de relever certaines incohérences, notamment dans les voies de synthèse de la méthionine et de l'ubiquinone (de Berardinis et al. 2008). Ces constats conduisirent à reconsidérer le fonctionnement de ces deux voies et appellent d'autres expériences pour en élucider précisément la structure.

De manière à déterminer l'essentialité des gènes sur d'autres environnements, l'équipe Thesaurus a mis en place une procédure de phénotypage de l'ensemble de la collection de mutants sur milieu liquide (de Berardinis et al. 2008). Cette procédure, également haut débit et basée sur des tests de croissance clonaux, se déroule entièrement sur plaques 96 puits. Après une préculture des mutants dans le milieu initial de sélection, ceux-ci sont inoculés dans des puits contenant le milieu à tester. Suivent 24 heures d'incubation au terme desquels la densité optique à 600 nm de chacun des puits est mesurée afin d'en quantifier la quantité de cellules présente. Cette procédure a été appliquée pour un panel relativement large d'environnements. Le phénotypage de la collection de mutants sur glucarate et galactarate a ainsi permis de caractériser de nouveaux gènes impliqués dans leur dégradation (Aghaie et al. 2008), tandis que les résultats de phénotypage sur 2,3-butanediol et quinate ont corroboré les voies de dégradation précédemment connues et permis de formuler des hypothèses sur de nouveaux gènes impliqués (de Berardinis et al. 2008).

5 Synthèse et objectifs de la thèse

En conclusion de cette partie introductive, nous soulignerons les points suivants :

- Le métabolisme global d'un organisme peut être en grande partie reconstruit à partir de l'annotation de son génome, ouvrant la voie à l'exploration à grande échelle des caractéristiques métaboliques des organismes séquencés. Cependant, les réseaux métaboliques reconstruits de cette manière sont fortement biaisés par les méthodes d'annotation, qui s'appuient pour l'essentiel sur les activités enzymatiques déjà identifiées ; ils sont très probablement incomplets.

- L'étude des phénotypes de croissance donne une perspective macroscopique sur le métabolisme des organismes (leur physiologie). Associée à des techniques d'inactivation génétique, elle permet d'explorer l'essentialité des gènes et d'associer certains d'entre eux à la survie sur des environnements particuliers. Ces résultats peuvent compléter utilement la connaissance issue de l'annotation. Néanmoins, leur bonne interprétation d'un point de vue du métabolisme nécessite dans de

nombreux cas de considérer le fonctionnement global du métabolisme tout en tenant compte des environnements de croissance.

- Les modèles métaboliques à base de contraintes se sont révélés bien adaptés à prédire les phénotypes de croissance à partir d'une description simple des réactions du métabolisme. Des travaux, récents au début de la thèse, ont montré que la confrontation des phénotypes prédits aux phénotypes observés permettait d'identifier des conflits entre la connaissance du réseau métabolique et ces observations expérimentales. Cependant, l'exploitation des phénotypes par les modèles manque d'un cadre d'interprétation clair et de méthodes permettant de guider les corrections à réaliser au réseau métabolique.

Le contexte scientifique du Genoscope s'est montré particulièrement favorable à l'entreprise d'une étude globale du métabolisme à l'aide de phénotypes de croissance. D'une part, l'annotation experte du génome d'*A. baylyi* offrait une base solide à la reconstruction complète de son métabolisme et, d'autre part, les capacités expérimentales de phénotypage de ses mutants fournissaient un ensemble de données d'essentialité sans précédent pour cette bactérie.

Nous avons donc décidé d'explorer plus en détail les méthodes d'interprétations des phénotypes de croissance à l'aide des modèles métaboliques, sur la base du métabolisme d'*A. baylyi*. Dans cette optique, notre thèse s'attacha à atteindre trois objectifs successifs. Tout d'abord, évaluer la performance des méthodes de l'état de l'art pour reconstruire un modèle métabolique global fiable à partir de l'annotation et de la connaissance initiale, et proposer des solutions aux faiblesses constatées. Ensuite, confronter les phénotypes prédits par le modèle aux phénotypes observés et proposer une méthodologie d'interprétation permettant d'exploiter au mieux leurs incohérences dans le but de corriger le modèle reconstruit. Enfin, élaborer une méthode originale de correction automatique des modèles à partir de données phénotypiques expérimentales. Les résultats de nos recherches sur ces trois objectifs sont présentés dans les trois grandes parties suivantes.

RECONSTRUCTION D'UN MODELE GLOBAL DU METABOLISME D'*ACINETOBACTER BAYLYI* ADP1

Dans cette première partie consacrée à nos résultats, nous exposerons le processus de reconstruction que nous avons mis en place pour obtenir un modèle global du métabolisme d'*A. baylyi* qui représente le plus fidèlement possible notre connaissance de son métabolisme. À cette occasion, nous présenterons les méthodes et vérifications que nous avons notamment élaborées afin d'adapter la représentation des réactions aux spécificités de la modélisation. Dans un deuxième temps, nous décrirons les caractéristiques marquantes du modèle obtenu.

6 Processus de reconstruction

Fin 2005, alors même que les méthodes de modélisation du métabolisme gagnaient en popularité, seules quelques équipes avaient entrepris de reconstruire des modèles globaux pour un nombre limité d'organismes. Il s'avérait en effet – et à juste titre – qu'une telle entreprise était une tâche laborieuse dont les difficultés n'étaient pas encore toutes clairement identifiées. Des initiatives commençaient à être mises en place pour répertorier et résoudre ces difficultés (Segrè et al. 2003), mais le processus de reconstruction ne semble se rapprocher qu'aujourd'hui de la maturité, comme peuvent en témoigner les apparitions récentes d'outils (DeJongh et al. 2007; Schwarz et al. 2007) et d'articles de revue (Durot et al. 2009; Feist et al. 2009; Reed, Famili et al. 2006). Pour cette raison, nous présenterons notre processus de reconstruction du modèle d'*A. baylyi* en soulignant au lecteur les points-clés liés aux exigences de la modélisation que nous aurons identifiés.

Deux phases quasiment indépendantes se distinguent dans le processus de reconstruction (voir Figure 22). La première consiste à répertorier les activités métaboliques connues de l'organisme, à partir de l'annotation de son génome mais également de la littérature biochimique et d'informations physiologiques. La seconde adapte la représentation de ces réactions afin de construire un modèle compatible avec les hypothèses de modélisation. Ces deux étapes sont décrites séparément dans les deux sections suivantes.

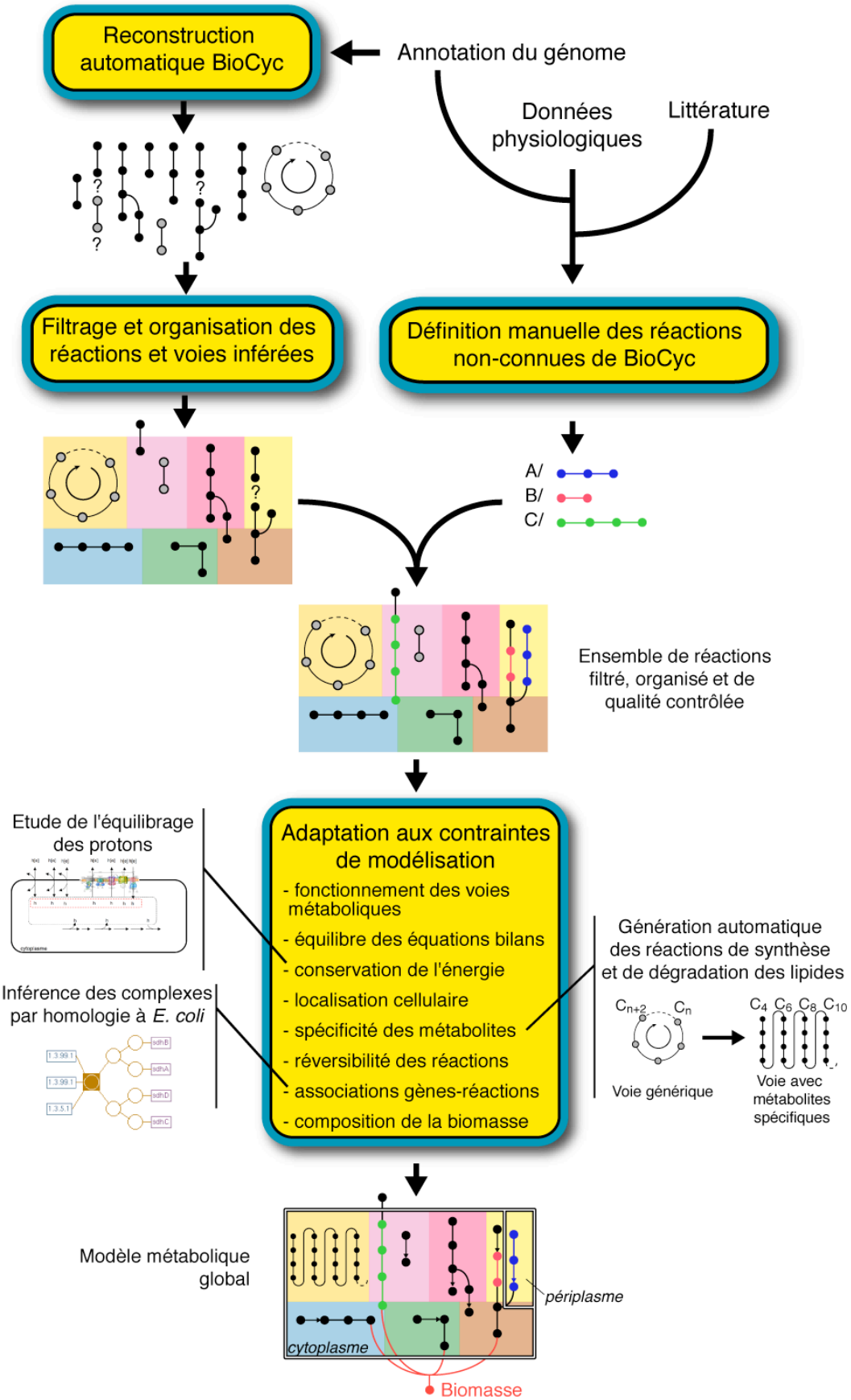


Figure 22. Processus de reconstruction du modèle métabolique d'*A. baylyi*. Le réseau métabolique est schématisé par un graphe. Les nœuds grisés indiquent des métabolites génériques, les arrière-plans colorés des grandes catégories fonctionnelles du métabolisme.

6.1 Identification des activités métaboliques

L'annotation experte du génome d'*A. baylyi*, réalisée avec une attention particulière pour les processus métaboliques, offre une base solide à la reconstruction de son réseau métabolique. Le travail des experts a en effet contribué non seulement à vérifier et valider les annotations assignées automatiquement, mais également à en préciser certaines et à en définir des nouvelles à partir des connaissances spécifiques à la bactérie. L'ensemble des annotations représente donc une source d'information relativement complète sur le métabolisme d'*A. baylyi,* sur laquelle nous avons choisi d'appuyer notre processus de reconstruction. De plus, contrairement aux quelques organismes modèles dont la biochimie a été largement étudiée expérimentalement, l'annotation du génome représentait pour *A. baylyi* la seule source d'information pour une large majorité de ses activités enzymatiques. Le processus de reconstruction suivi est donc transposable à tout nouvel organisme dont le génome est séquencé et annoté.

La première étape du processus a consisté à traduire les annotations textuelles du génome en réactions définies par leurs équations bilans (voir Figure 22). Pour cela, nous nous sommes reposés sur les outils de l'état de l'art. Deux principales ressources étaient disponibles pour réaliser cette opération à grande échelle, à savoir les bases de données métaboliques KEGG (Kanehisa et al. 2004) et BioCyc (Karp et al. 2002).

Nous avons rapidement arrêté notre choix sur BioCyc. Bien que répertoriant un vaste ensemble de réactions dans ses bases de données, KEGG ne disposait pas de méthodes aussi élaborées que celles de BioCyc pour les associer automatiquement aux annotations des génomes. En effet, KEGG se repose pour cela exclusivement sur les identifiants EC attribués dans les annotations et est donc sensible à la qualité de ces attributions. Dans la pratique, une proportion significative des gènes de fonction enzymatique n'est pas annotée avec un identifiant EC complet, rendant leur interprétation par KEGG impossible. Par exemple pour *A. baylyi*, pour 794 gènes codant pour des enzymes, seuls 576 sont annotés avec un identifiant EC complet et 103 avec un identifiant incomplet, laissant 115 gènes sans identifiant EC[45]. De plus, à ces absences d'annotation EC, causées principalement par la non-exhaustivité de la

[45] Décompte réalisé à partir de l'annotation d'*A. baylyi* disponible dans MaGe au 28 avril 2009 (http://www.genoscope.cns.fr/agc/mage/). Le même décompte réalisé pour *E. coli* K12 donne des proportions similaires.

classification (toutes les activités enzymatiques n'y sont pas représentées) et les difficultés des annotateurs à déterminer sans ambiguïté les bons identifiants, s'ajoutent des problèmes de compatibilité entre identifiants provoqués par les modifications périodiques de la classification.

Contrairement à KEGG, BioCyc exploite, en complément des identifiants EC, les noms d'enzymes attribués textuellement dans l'annotation, rendant la détection des réactions plus performante (Karp et al. 2002). En outre, son logiciel de reconstruction – Pathway Tools – réalise la reconstruction voie métabolique par voie métabolique, permettant ainsi d'inférer d'éventuelles réactions manquantes (voir Introduction, sections 1.4.2 et 3.2.1).

Nous avons exécuté Pathway Tools (version 8.0) sur l'annotation d'*A. baylyi* et obtenu en résultat la base de donnée AcinetoCyc. Cette dernière contient l'ensemble des réactions métaboliques identifiées par Pathway Tools pour *A. baylyi*. Chacune d'entre elles y est décrite par son équation chimique et est reliée aux gènes et enzymes la catalysant. AcinetoCyc contient ainsi l'information nécessaire à la construction d'un modèle métabolique.

Le processus de reconstruction automatique BioCyc présente cependant certains inconvénients rendant impossible l'utilisation directe des réactions inférées. Tout d'abord, Pathway Tools tend à « surprédire » les voies métaboliques. En effet, il suffit parfois qu'une petite fraction des enzymes d'une voie soit codée dans le génome pour que l'ensemble de la voie et toutes ses réactions soient prédites par Pathway Tools. Cette méthode engendre ainsi un nombre conséquent de faux positifs – des voies métaboliques et réactions sans gène associé n'ayant pas de réalité biologique. Ensuite, le procédé de détection des réactions à partir des annotations textuelles commet parfois des erreurs et infère des réactions ne correspondant pas à l'annotation ; ce cas de figure apparaît notamment lorsque la spécificité des substrats n'est pas précisée dans l'annotation (ex. : alcool déshydrogénase). De plus, certaines activités enzymatiques ne sont pas encore répertoriées dans la base de données de référence de BioCyc (MetaCyc) ; lorsqu'elles sont présentes dans une annotation, Pathway Tools ne peut créer les réactions correspondantes. Enfin, bien que l'ensemble des réactions inférées soit organisé en voies métaboliques, ces dernières sont nombreuses, souvent isolées les unes des autres et parfois redondantes. Il est difficile de ce fait

d'appréhender la structure globale du réseau métabolique et de comprendre les rôles respectifs des voies métaboliques.

Pour ces raisons et afin de garantir la qualité du réseau métabolique reconstruit, le processus automatique de création des réactions devait nécessairement être complété par une phase de curation manuelle. De manière à maîtriser au mieux le contenu du modèle et étant donné le grand nombre de « faux positifs » générés par Pathway Tools, nous avons choisi de construire le modèle progressivement en ajoutant une à une les réactions le constituant. Nous avons pour cela défini sept grandes catégories métaboliques dans lesquelles nous avons réparti les réactions du réseau global :

- Métabolisme central,
- Métabolisme des acides aminés
- Métabolisme des lipides et des constituants de la membrane
- Métabolisme des nucléotides
- Biosynthèse des cofacteurs
- Voies de dégradations
- Processus de transport

Pour chaque catégorie fonctionnelle, nous avons examiné les voies métaboliques concernées prédites dans AcinetoCyc. Nous avons alors retenu dans le modèle uniquement les réactions pour lesquelles suffisamment d'informations justifiaient leur présence : bonne adéquation de l'annotation du gène à la réaction, importance de la réaction dans la catégorie fonctionnelle, participation de la réaction à une voie métabolique connue de l'organisme. Ce processus de sélection des réactions fut réalisé en utilisant systématiquement les connaissances initiales du métabolisme d'*A. baylyi* contenues dans la littérature. Nous avons ainsi pu d'une part valider la présence de réactions sans gène associé et d'autre part compléter la reconstruction en ajoutant les réactions connues qui n'avaient pas été inférées par Pathway Tools (voir Figure 22). Ce dernier cas comprend (1) les voies métaboliques spécifiques à *A. baylyi* décrites dans la littérature, (2) les processus de transport de métabolites mis en évidence grâce aux connaissances sur la physiologie de la bactérie et aux prédictions de la ressource TransportDB (Ren et al. 2004), et (3) les activités enzymatiques présentes dans l'annotation mais n'ayant pu être interprétées par Pathway Tools. De plus, comme nous le verrons dans la partie suivante, le bon fonctionnement du modèle

nécessite d'avoir des voies métaboliques complètes, sans réaction manquante. Nous avons de ce fait réalisé ces ajouts de réactions dans le souci d'obtenir des voies métaboliques complètes et fonctionnelles.

Cette étape de curation manuelle – basée notamment sur l'examen de 66 références bibliographiques (voir partie 7.3) – nous a amené à compléter significativement le réseau métabolique, comme en témoigne la Figure 23. Les catégories fonctionnelles incluant le plus de réactions ajoutées manuellement sont celles des voies de dégradation et des processus de transport. La première comprend en effet un grand nombre d'activités relativement spécifiques à *A. baylyi* et non inférées dans AcinetoCyc. Les processus de transport ne sont quant à eux pas prédits par la méthode standard de Pathway Tools et ne sont de ce fait pas inclus dans AcinetoCyc.

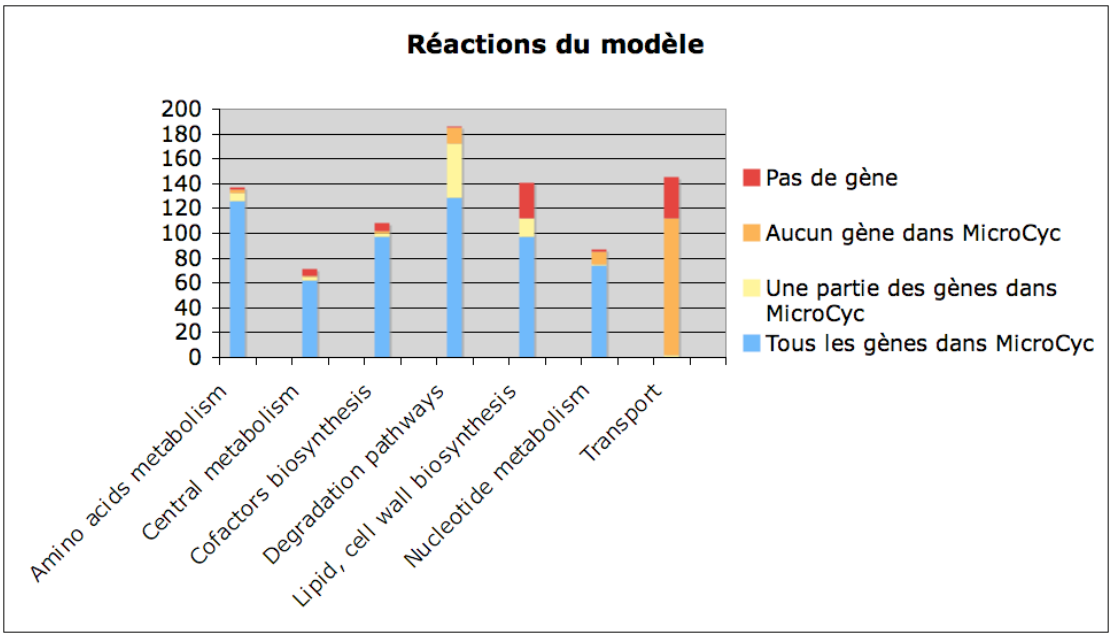


Figure 23. Proportion de réactions du modèle représentées dans MicroCyc. MicroCyc contient les reconstructions automatiques du métabolisme réalisées périodiquement par Pathway Tools à partir des génomes contenus dans MicroScope (http://www.genoscope.cns.fr/agc/microcyc). Compte tenu de l'évolution des identifiants de réactions BioCyc, nous avons choisi d'effectuer la comparaison entre le modèle et la reconstruction MicroCyc sur la base des gènes enzymatiques pris en compte par MicroCyc. La reconstruction MicroCyc utilisée ici a été effectuée par Pathway Tools version 13.0 sur l'annotation d'*A. baylyi* d'avril 2009.

Outre le fait de fournir une reconstruction organisée du métabolisme, le processus de curation systématique de chaque réaction offre la possibilité de gérer la fiabilité de leur présence, ce qui n'est pas effectué dans la reconstruction automatique de Pathway Tools. Nous avons ainsi associé à chaque réaction un score de confiance évaluant la

fiabilité des informations soutenant sa présence. Celui-ci est principalement basé sur le score de confiance attribué aux annotations de ses gènes et prend en compte les informations supplémentaires issues de la littérature :

- 1 : activité confirmée expérimentalement dans *A. baylyi*
- 2 : gène annoté grâce à une homologie forte à un gène de fonction prouvée, et activité définie précisément dans l'annotation
- 3 : annotation du gène réalisé à partir d'indice de moindre confiance
- 4 : présence de la réaction inférée pour compléter des voies métaboliques essentielles.

Ces scores permettent d'évaluer la qualité de la reconstruction et pourront ensuite être utilisés pour guider des corrections.

En conclusion, bien que significativement accélérée par les outils d'exploitation automatique des annotations, l'énumération des réactions métaboliques nécessite un travail de curation significatif, ne serait-ce que pour intégrer les connaissances biochimiques non représentées dans les annotations.

D'un point de vue pratique, nous avons effectué ce travail au sein d'un tableur regroupant les informations sur les réactions. En effet, aucun logiciel spécialisé n'offrait la souplesse nécessaire à la construction d'un modèle de cette taille par ce processus[46]. Les bases de données métabolique BioCyc, et a fortiori AcinetoCyc, offre la possibilité d'être modifiée manuellement, mais elles se révèlent difficilement utilisables pour manipuler simultanément de grands ensembles de réactions. Nous avons donc exporté les réactions d'AcinetoCyc afin de les traiter dans le tableur. Nous avons toutefois retranscrit dans AcinetoCyc une partie des modifications effectuées lors de la construction du modèle – nous y avons notamment ajouté les voies métaboliques supplémentaires – pour pouvoir utiliser AcinetoCyc à des fins de visualisation[47].

[46] Les logiciels classiques de construction des modèles métaboliques (p.ex. CellDesigner (Funahashi et al. 2003)) ne sont en effet adaptés qu'à des réseaux de taille modeste.
[47] AcinetoCyc est accessible à l'adresse http://www.genoscope.cns.fr/acinetocyc/

6.2 Adaptation aux « contraintes » de modélisation

Comme déjà évoqué dans notre revue sur les modèles à base de contraintes (voir 3.2.1), leur construction nécessite de prendre en compte certaines exigences liées aux hypothèses de modélisation. C'est pourquoi, dans la deuxième phase du processus de reconstruction (voir Figure 22), nous avons (1) identifié ces exigences, (2) vérifié que le modèle y répondait et, le cas échéant, (3) appliqué des modifications pour adapter le modèle. Dans cette partie, nous exposerons l'ensemble des vérifications et modifications que nous avons effectué pour le modèle d'*A. baylyi* et détaillerons nos contributions les plus significatives par rapport à l'état de l'art.

6.2.1 Fonctionnement des voies métaboliques

La contrainte de conservation de la masse impose aux flux de production et de consommation de chaque métabolite interne de s'équilibrer au sein du modèle. Pour être valide, cette contrainte suppose donc que toutes les réactions majoritairement impliquées dans ces conversions sont incluses dans le modèle. Si tel n'était pas le cas, cette contrainte imposerait des liens erronés entre les flux de réactions. Pour cette raison, il est nécessaire d'inclure dans le modèle toutes les réactions majeures impliquant chacun de ses métabolites, dans la limite, bien entendu, des activités enzymatiques identifiées dans l'organisme.

Cette vérification est particulièrement importante car l'absence d'une réaction peut provoquer le « blocage » d'une voie métabolique, voire du modèle entier. En effet, si cette réaction est par exemple indispensable à la production d'un métabolite, la contrainte de conservation de la masse imposera aux réactions consommant ce métabolite un flux nul, lui-même potentiellement propagé à d'autres réactions par cette même contrainte. Dans notre processus de reconstruction, nous avons donc prêté une attention particulière à inclure des voies métaboliques complètes dans le modèle et à vérifier la présence pour chaque métabolite de réactions de consommation et de production. Cette vérification semble naturelle pour les métabolites intermédiaires de grandes voies métaboliques linéaires, elle l'est moins, mais tout autant nécessaire, pour les cofacteurs et autres métabolites dont les processus de production et consommation sont répartis entre les catégories fonctionnelles du métabolisme.

Cette vérification nous a conduit dans quelques cas à inférer et ajouter au modèle des réactions indispensables au fonctionnement des voies métaboliques. Ces réactions ont été choisies en examinant les voies métaboliques présentes dans *A. baylyi* et déterminant les réactions les plus probables pour combler les conversions métaboliques absentes. Dans la majorité des cas, nous nous sommes appuyés pour cela sur les voies métaboliques connues chez les autres organismes et automatiquement inférées par Pathway Tools (voir Figure 24). Pour refléter le peu d'indices confirmant la présence de ces réactions, nous leur avons attribué un score de fiabilité faible. Dans un rare cas, nous n'avons pu déterminer de réaction consommant un métabolite : il s'agit du s-adenosyl-4-methylthio-2-oxobutanoate produit lors de la biosynthèse de la biotine. Pour permettre le fonctionnement de la voie, nous avons ajouté une réaction d'échange supplémentaire consommant artificiellement ce métabolite.

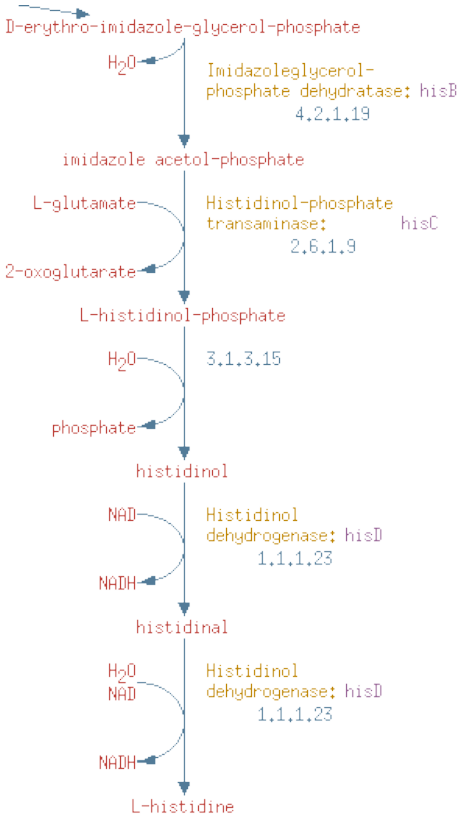


Figure 24. Dernières étapes de la voie de biosynthèse de l'histidine. La réaction manquante (histidinol phosphatase, EC 3.1.3.15) a été inférée dans le modèle pour permettre le fonctionnement de la voie métabolique. Sa présence est cependant suggérée par l'existence des autres réactions de la voie et la présence dans le génome de gènes annotés comme phosphatases (sans mention du substrat). Illustration issue de AcinetoCyc.

Malgré ces vérifications, la contrainte de conservation de la masse peut tout de même provoquer des « blocages » plus complexes à interpréter. Pour faciliter leur

détection, nous avons vérifié et validé le « fonctionnement » des voies métaboliques au fur et à mesure de leur introduction dans le modèle. Cette stratégie nous a dispensé d'utiliser les méthodes dédiées à la détection de ces blocages (voir revue section 3.2.1), dont les versions réellement performantes ne sont apparues que tardivement (Kumar et al. 2007; Senger & Papoutsakis 2008).

6.2.2 Équilibre des équations bilans

La bonne application de la contrainte de conservation de la masse exige également que les équations bilans des réactions soient correctement équilibrées. L'absence d'un substrat ou d'un produit dans une équation fausse le bilan global de la réaction et peut provoquer la production et consommation artificielles de certains métabolites. Seules les réactions d'échanges, dont l'objectif est justement d'introduire ou d'extraire des métabolites dans le modèle, ne sont pas équilibrées.

Nous nous sommes basés sur les formules chimiques des métabolites pour vérifier l'équilibre des équations bilans, élément par élément. Toutefois, les métabolites pouvant se trouver sous différents états de protonation, nous n'avons pas vérifié l'équilibre des équations au proton près, en dehors des réactions impliquant le transport de protons à travers la membrane cytoplasmique (voir partie suivante). Pour chaque réaction à l'équation non équilibrée, nous avons alors cherché à résoudre l'incohérence à l'aide d'autres ressources sur leur biochimie, dont principalement BRENDA (Barthelmes et al. 2007). Les réactions créées dans AcinetoCyc contenaient une proportion non négligeable d'erreurs dans leurs équations. Les dernières versions de BioCyc ont cependant corrigé la grande majorité d'entre elles[48].

Il est important de noter également que le fait d'équilibrer globalement toutes les réactions du modèle implique de connaître spécifiquement tous les métabolites. Les réactions définies à l'aide de métabolites génériques (par exemple *un acide dicarboxylique*, représentant des acides carboxyliques de diverses chaînes carbonées) ne peuvent ainsi être reliées aux autres réactions du modèle, alors même qu'elles sont correctement équilibrées. Nous verrons plus loin au paragraphe 6.2.5 les méthodes que nous avons mises en œuvre pour y répondre.

[48] Voir l'historique des améliorations à l'adresse http://metacyc.org/release-notes.shtml .

6.2.3 Conservation de l'énergie

Afin de réaliser des conversions métaboliques à la thermodynamique peu favorable, certaines réactions se couplent à des processus leur apportant de l'énergie (voir en introduction section 1.2.6). Ces derniers sont majoritairement de nature chimique – conversion exothermique d'un métabolite très énergétique, par exemple l'ATP, en un métabolite moins énergétique, l'ADP – mais également physique – utilisation d'un gradient de concentration de part et d'autre d'une membrane cellulaire. Lorsque les échanges d'énergie se limitent à ces deux formes – ce qui est majoritairement le cas pour les organismes qui exploitent l'énergie chimique contenue dans les métabolites du milieu – la contrainte de conservation de la masse permet de les prendre en compte dans le modèle. En appliquant cette contrainte aux métabolites sous leurs différentes formes énergétiques, elle impose aux flux responsables de leur production de compenser leur consommation.

Ainsi, dans l'exemple de l'ATP, la contrainte de conservation de la masse impose aux réactions régénérant l'ATP de le produire à un flux comblant la consommation d'ATP par les réactions du métabolisme. À leur tour, les réactions régénérant l'ATP imposent un flux aux réactions leur permettant d'avoir lieu (cycle de Krebs), permettant *in fine* de tenir compte de la demande énergétique dans le fonctionnement global du métabolisme.

La contrainte de conservation de la masse modélise de manière similaire les échanges d'énergie liés aux gradients de concentration. Dans ce cas, le niveau d'énergie des métabolites est déterminé par leur localisation, interne ou externe à la membrane cellulaire. Le modèle représente séparément les métabolites localisés différemment, et traduit les processus de transport par des réactions « déplaçant » ces métabolites d'un compartiment à un autre. En imposant la conservation de la quantité des métabolites internes, le modèle impose alors indirectement un équilibre des échanges au travers de la membrane forçant le maintien du gradient de concentration dans le modèle (voir Figure 25**A**).

L'utilisation d'un gradient de concentration comme vecteur énergétique s'applique principalement aux protons : certaines réactions entretiennent ce gradient en expulsant des protons (notamment les réactions de la chaîne respiratoire) tandis que d'autres exploitent son énergie en laissant entrer des protons (notamment l'ATP synthase et

des processus actifs de transport). La difficulté à déterminer les états de protonation des métabolites complique cependant l'obtention de réactions équilibrées au proton près et, par conséquent, l'application de la contrainte de conservation de la masse pour les protons intracellulaires. Nous avons choisi de contourner cette difficulté en supposant que seuls les processus de transport de protons contribuaient significativement à leurs conversions. Nous n'avons donc équilibré au proton près que les réactions transportant les protons, et appliqué la contrainte de conservation de la masse uniquement aux protons transportés (voir Figure 25**B**).

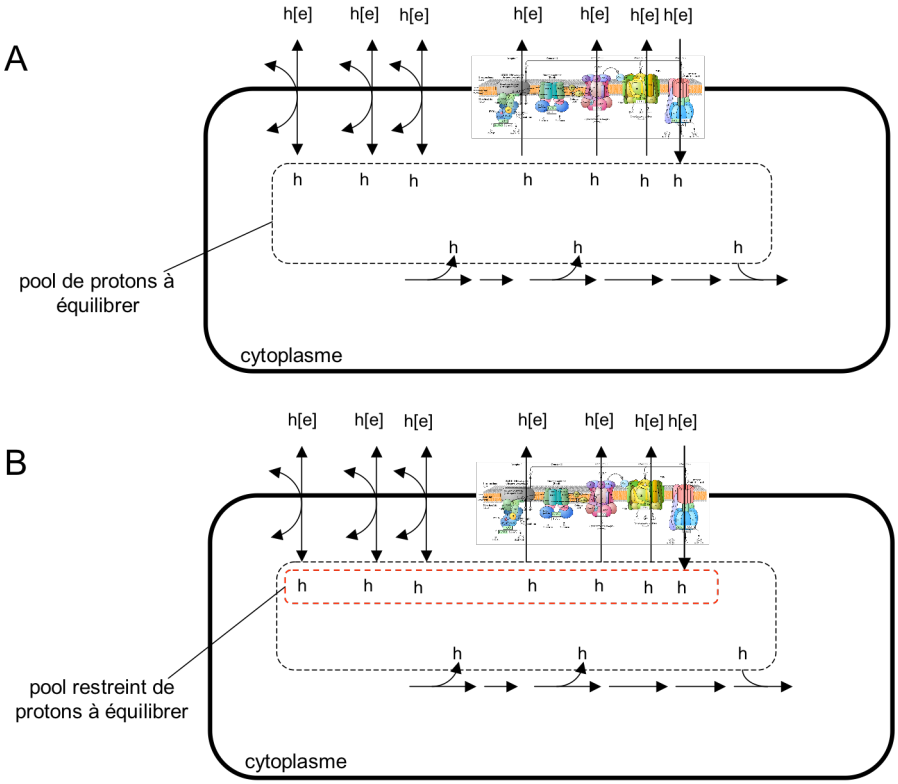


Figure 25. Contrainte de conservation de la masse appliquée aux protons. A. La contrainte est appliquée à l'ensemble des protons intracellulaires et exige alors d'équilibrer au proton près toutes les équations bilans. **B.** La contrainte est appliquée uniquement aux processus transportant les protons, ces derniers étant supposés contribuer majoritairement aux conversions de proton.

Afin d'évaluer les effets de ce choix de modélisation sur les prédictions du modèle, nous avons comparé le modèle d'*E. coli* iJR904 équilibrant tous les protons intracellulaires (Reed et al. 2003) à une version que nous avons modifié pour n'équilibrer que les protons des processus de transport (voir Figure 25).

Nous avons tout d'abord déterminé les taux de croissance prédits pour chacun de ces modèles sur un ensemble de 10 000 conditions environnementales composées d'un ensemble minimal fixe de molécules (dioxyde de carbone, eau, proton, fer II,

potassium et sodium) complété aléatoirement par un métabolite de chacune des catégories suivantes : accepteur d'électron (4 métabolites), sources de carbone (89 métabolites), sources d'azote (34 métabolites), sources de phosphore (4 métabolites) et sources de souffre (2 métabolites). Pour chaque environnement, nous avons prédit le taux de croissance optimal de chaque modèle par la méthode Flux Balance Analysis (Varma & Palsson 1994) et comparé leurs valeurs. La Figure 26 trace l'histogramme des différences relatives entre taux de croissance, calculé par la formule :

$$D_{rel} = \frac{\mu_{\text{modèle simplifié}} - \mu_{\text{modèle complet}}}{\max\left(\mu_{\text{modèle simplifié}}, \mu_{\text{modèle complet}}\right)}$$

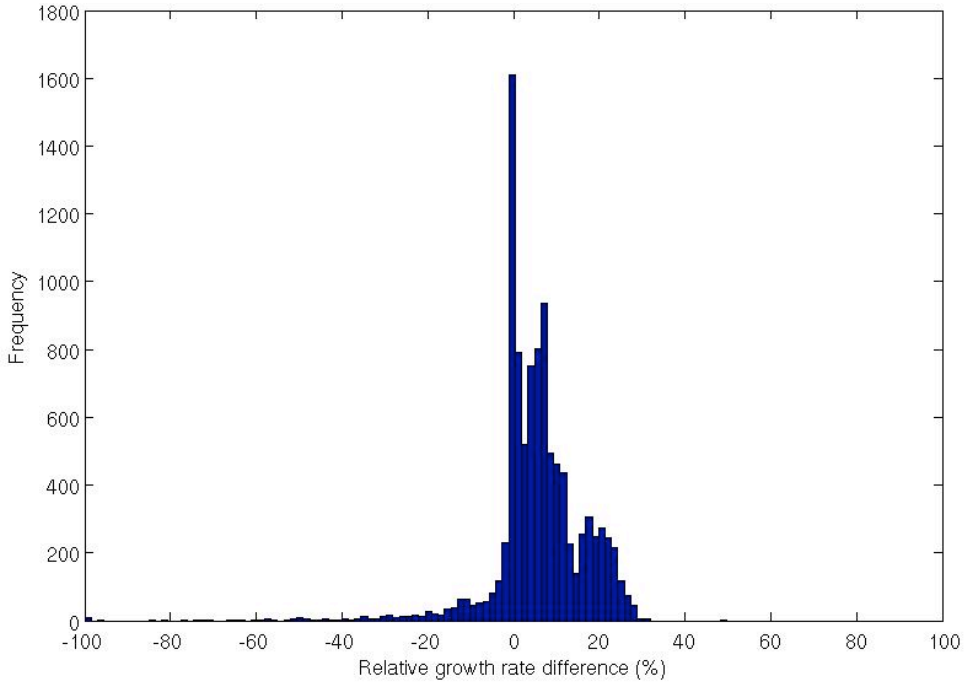


Figure 26. Histogramme des différences relatives entre les taux de croissance prédits par les deux modèles sur les 10000 environnements simulés.

Dans 99% des environnements, la différence relative est contenue entre -30% et 30%. Pour une majorité d'entre eux (75%), le modèle simplifié prédit un taux de croissance supérieur à celui du modèle complet. Ce résultat est dû au fait que, en moyenne, les voies métaboliques requises pour la croissance produisent plus de protons qu'elles n'en consomment. Le modèle complet tient compte de ces protons, ce qui pénalise légèrement l'équilibre du gradient. Dans ce modèle, plus d'énergie est consacrée au maintien du gradient de concentration.

Pour les 1% d'environnements restants, le modèle simplifié prédit au contraire un taux de croissance significativement plus faible que celui du modèle complet (diminution supérieure à 30%). Pour 10 environnements de ce type, le modèle réduit prédit un taux de croissance nul alors même que le modèle complet prédit un taux de valeur classique. L'examen des voies métaboliques utilisées pour croître sur ces environnements a révélé des voies pouvant capturer des protons internes et les excréter hors de la cellule au sein d'autres métabolites. Par exemple, la voie représentée sur la Figure 27 importe de la guanine de l'environnement, la convertit en ammonium et xanthine qui sont ensuite excrétés dans l'environnement. Dans le modèle, la guanine deaminase (***GUAD***) capture un proton qui est ensuite excrété sous la forme de l'ion ammonium. Le bilan net de cette voie consomme un proton intracellulaire (voir Figure 27). En utilisant cette voie avec un flux élevé, le modèle complet réussit à maintenir une excrétion de proton suffisante à la croissance.

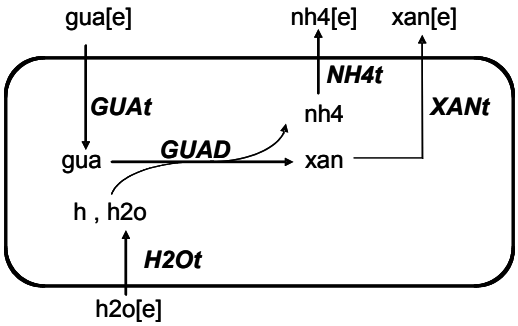


Figure 27. Consommation nette de proton intracellulaire par deamination de la guanine. Le bilan net de cette voie est en effet gua[e] + h2o[e] + **h** → nh4[e] + xan[e]. gua, guanine; xan, xanthine ; nh4, ammonium ; h, proton ; h2o, eau ; ***GUAD***, guanine deaminase.

Les différences majeures de prédictions de taux de croissance correspondent à des situations similaires impliquant des voies de ce type. Ces voies ne correspondent cependant pas à leur utilisation physiologique réelle, le flux nécessaire pour exporter les protons de cette manière étant sans comparaison avec leurs flux habituels. Bien que, dans l'ensemble, le modèle complet tienne compte correctement de l'effet des consommations internes de proton, certains cas de figures semblent ainsi fausser la prise en compte des processus majeurs de maintien du gradient.

Dans un deuxième temps, nous avons évalué l'effet du type de modélisation sur la prédiction de phénotypes de croissance de mutants. Pour chacun des 10 000 environnements, nous avons prédit le taux de croissance d'un mutant de délétion simple de gène choisi au hasard, et calculé la diminution relative au taux de

croissance prédit pour la souche sauvage. La Figure 28 présente les différences de prédiction de ces diminutions relatives prédites par les deux types de modélisation.

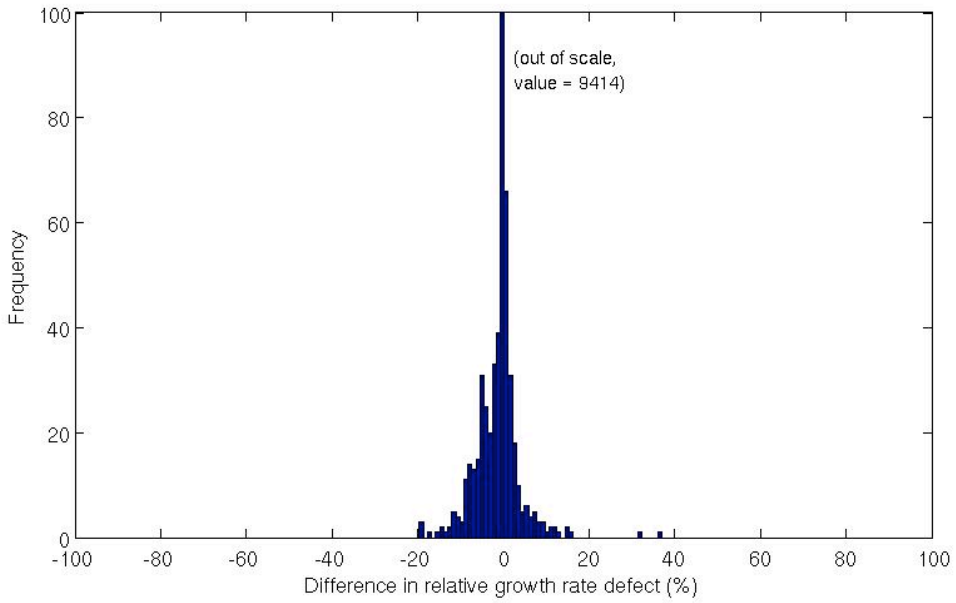


Figure 28. Histogramme des différences entre les diminutions relatives de taux de croissance par délétion de gène prédites par les deux modèles sur les 10000 environnements simulés.

Dans la très grande majorité des cas (94%), la diminution de croissance est identique. Ce résultat inclut cependant un grand nombre de cas (la moitié) pour lesquels la délétion est sans effet. Dans 2% des cas, la souche sauvage ne pouvait croître, empêchant le calcul de la diminution de croissance. Seuls 4% des cas présentaient une différence de prédiction de diminution de croissance entre les types de modélisation, de valeur relativement limitée (la majorité est contenue entre -20% et +20%).

En conclusion, cette étude nous a montré que la modélisation de l'équilibre des protons restreinte aux processus de transport (1) prédit des taux de croissance similaires à ceux d'une modélisation complète (quoique supérieur en moyenne), (2) prédit des phénotypes de croissance de mutants très similaires à ceux de la modélisation complète et (3) évite l'occurrence de voies métaboliques modifiant l'équilibre des protons de façon physiologiquement irréaliste. Nous avons donc choisi d'appliquer ce type de modélisation au modèle d'*A. baylyi*.

La stœchiométrie des réactions transportant les protons à travers la membrane est un autre point nécessitant une attention particulière du modélisateur. Leur valeur peut

en effet influer significativement sur l'efficacité du métabolisme à générer de l'énergie. Cependant, ces stœchiométries peuvent varier en fonction du processus de transport, des conditions environnementales (pH) et du gradient de concentration ; leurs valeurs ne sont pas fixées de manière aussi rigoureuse que celles des conversions chimiques. Des études pour quelques organismes modèles ont permis d'en déterminer des valeurs moyennes pour des conditions environnementales précises (Gennis & Stewart 1996) ; dans le cas d'*A. baylyi*, nous avons retenu des valeurs similaires à celles de ces organismes car aucune étude précise n'avait été réalisée pour *A. baylyi*. Comme évoqué dans notre revue sur les modèles (section 3.2.1), ces valeurs peuvent également être inférées à partir de données physiologiques à l'aide des modèles.

6.2.4 Localisation cellulaire

La contrainte de conservation de la masse ne s'applique en toute rigueur qu'aux pools de métabolites situés dans le même compartiment cellulaire. Deux enzymes situées dans deux compartiments différents n'opèrent en effet pas sur le même ensemble de métabolites. Il est donc essentiel de tenir compte de la localisation des enzymes pour appliquer correctement cette contrainte.

Dans le cas d'*A. baylyi*, bactérie gram-négative, seuls deux compartiments segmentent son métabolisme : le cytoplasme et le périplasme. La très grande majorité de son métabolisme se déroule dans le cytoplasme, seules quelques voies bien identifiées opèrent dans le périplasme. Il s'agit notamment des premières étapes de la glycolyse et des voies de dégradations du quinate, du shikimate et du chlorogenate (Young et al. 2005). L'examen de la littérature associée à ces voies nous a permis de déterminer précisément la localisation des enzymes impliquées dans ces voies.

Les processus de transport doivent également être modélisés avec attention pour autoriser les échanges de métabolites entre compartiments. Lorsque l'annotation n'indiquait pas d'enzyme impliquée dans le transport d'un métabolite, nous nous sommes basés sur des informations physiologiques de l'organisme (métabolites connus pour être transportés) pour inférer ces réactions et compléter le modèle.

6.2.5 Spécificité des métabolites

Comme nous l'avons vu plus haut, l'équilibre des équations bilan appliqué à l'ensemble du réseau métabolique exige que chaque métabolite soit connu précisément. Cependant, les substrats précis de certaines enzymes ne sont pas toujours spécifiés dans l'annotation, soit parce qu'ils sont inconnus, soit parce que l'enzyme possède un large spectre de substrats. De même, pour représenter cette large spécificité, les réactions inférées par Pathway Tools utilisent des métabolites génériques, non compatibles avec la modélisation.

Pour chaque enzyme de ce type, nous avons cherché à déterminer les substrats spécifiques les plus probables. Nous nous sommes pour cela appuyés sur deux grandes sources d'information.

Nous avons tout d'abord exploré la littérature associée à ces enzymes et les bases de données enzymatiques (principalement BRENDA (Barthelmes et al. 2007)) pour rechercher des informations sur la caractérisation expérimentale de ces enzymes. Une large proportion de ces enzymes appartient aux voies cataboliques ; les études les ayant identifiées ont de ce fait souvent cherché à délimiter expérimentalement le spectre de substrats utilisables. Cette information n'est néanmoins en général pas reprise exhaustivement dans l'annotation et subsiste uniquement dans la littérature et les bases de données dédiées (BRENDA).

Nous avons également utilisé le contexte métabolique – constitué par les voies métaboliques déjà reconstruites – pour identifier les substrats potentiels jouant déjà un rôle dans le réseau métabolique et les plus à même de conférer à l'enzyme un rôle significatif dans le métabolisme. Les bases de données répertoriant les métabolites par catégories chimiques peuvent aider à énumérer tous les substrats potentiels (Degtyarenko et al. 2008; Fahy et al. 2009).

Certaines enzymes ont une spécificité particulièrement large et agissent sur un grand ensemble de métabolites. C'est le cas notamment des enzymes ayant pour substrats des molécules à longues chaînes carbonées, incluant de nombreux lipides. Les activités de ces enzymes sont alors représentées de manière synthétique à l'aide de métabolites génériques spécifiant les groupes fonctionnels chimiques mais laissant indéterminées les chaînes carbonées (voir Figure 29).

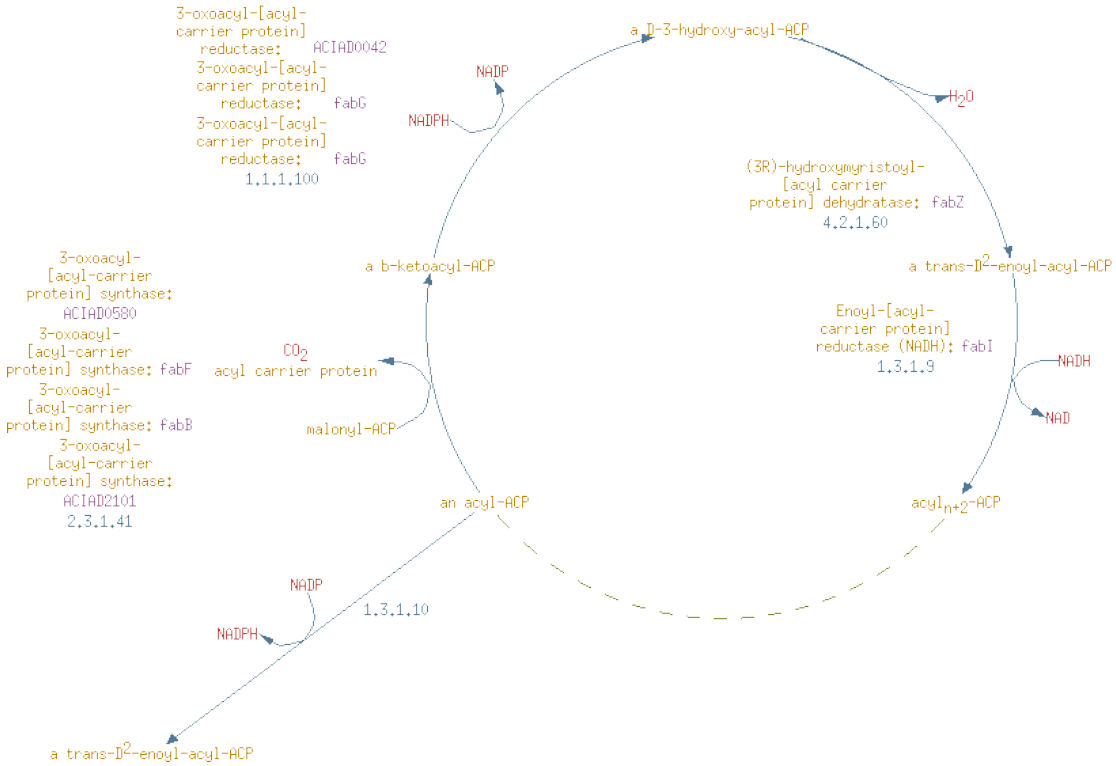


Figure 29. Voie d'élongation des acides gras représentée dans AcinetoCyc. L'élongation de la chaîne des acides gras est représentée à l'aide de métabolites génériques (an acyl-ACP, $acyl_{n+2}$-ACP) correspondant à des chaînes carbonées de longueurs arbitraires.

Afin de faciliter le travail de transformation de ces voies génériques en réactions aux métabolites spécifiques, nous avons développé un programme spécifiant automatiquement ces réactions pour des chaînes carbonées données en paramètre. Nous avons pour cela représenté chaque métabolite générique par une association entre un groupe fonctionnel et une chaîne carbonée de nature indéterminée. Nous avons ensuite formalisé les réactions génériques par des équations bilan impliquant ces métabolites génériques et indiquant formellement toute modification de la chaîne carbonée. Par exemple, la réaction générique correspondant à la 3-oxoacyl-ACP synthase est représentée sous la forme :

$$\text{malonyl-ACP} + \text{Cxx-ACP} \rightarrow \text{ACP} + CO_2 + \text{Cxx[+2]-3oxo-ACP}$$

où les métabolites Cxx-… sont les métabolites génériques et la notation [+2] indique un allongement de 2 carbones de la chaîne carbonée.

Les métabolites spécifiques furent ensuite dérivés des métabolites génériques en explicitant les chaînes carbonées. Nous avons caractérisé ces dernières par leurs longueurs (nombre de carbones) et les nombres, types (cis ou trans) et positions de leurs insaturations. De cette manière, une variété relativement large de chaînes

carbonées linéaires – correspondant à celles rencontrées majoritairement chez *A. baylyi* – pouvaient être prises en compte. Les réactions génériques furent alors explicitées pour chacune des chaînes carbonées considérées, en leur appliquant les éventuelles modifications de longueurs. Reprenant l'exemple de la 3-oxoacyl-ACP synthase pour une chaîne carbonée de longueur 12 possédant une insaturation de type cis entre le 3^{e} et le 4^{e} carbone (chaîne notée C12:1(c3)), le processus donne la réaction spécifique suivante :

malonyl-ACP + C12:1(c3)-ACP → ACP + CO_2 + C14:1(c5)-3oxo-ACP

Dans un deuxième temps, nous avons remplacé les noms systématiques générés pour les métabolites spécifiques par leurs identifiants dans AcinetoCyc et le modèle, lorsque ceux-ci existent. Cette étape est indispensable pour assurer la compatibilité des réactions créées avec le reste du modèle, et permettre leur « branchement » correct sur les autres voies métaboliques du réseau.

Nous avons utilisé ce programme pour générer les réactions spécifiques correspondant à trois voies métaboliques d'*A. baylyi* : l'élongation des acides gras, l'oxydation des acides gras et l'oxydation des acides dicarboxyliques. L'élongation des acides gras comprend cinq réactions génériques que nous avons implémentées pour les chaînes carbonées saturées paires de longueurs 2 à 18, et des chaînes paires de longueurs 12 à 18 possédant une insaturation, représentant un total de 55 réactions. Six réactions génériques composent la voie d'oxydation des acides gras, que nous avons implémentées pour les mêmes chaînes carbonées, représentant un ensemble de 66 réactions. Enfin, les cinq réactions génériques de la voie d'oxydation des acides dicarboxyliques furent spécifiées pour les chaînes carbonées saturées de longueurs 10 (sebacate), 8 (suberate), 6 (adipate), 7 (pimelate) et 5 (glutarate), constituant au total 25 réactions supplémentaires.

L'ensemble de ces réactions constitue une fraction significative du modèle métabolique global. L'utilisation d'un programme les générant automatiquement à partir des réactions génériques permet ainsi d'accélérer significativement la reconstruction du modèle. De plus, toute modification effectuée sur les réactions génériques peut de cette manière être propagée directement sur l'ensemble des réactions spécifiques. En effet, les réactions spécifiques dérivant d'une même réaction

générique partagent des caractéristiques *a priori* communes : réversibilité, stœchiométrie, lien avec les enzymes catalysant leur activité. Leur modification doit de ce fait être répercutée sur l'ensemble des réactions spécifiques, ce qui est rendu possible par leur génération automatique.

En conclusion, l'utilisation d'un programme générant automatiquement les réactions spécifiques à partir de réactions génériques et d'un ensemble de substrats semble nécessaire à la reconstruction rigoureuse et rapide des modèles. Il permet de gérer au mieux la nécessité de spécificité des modèles en séparant d'un coté la définition des activités enzymatiques (via des équations bilan génériques) et de l'autre l'énumération des substrats spécifiques réellement concernés.

6.2.6 Réversibilité des réactions

Les contraintes de réversibilité jouent un rôle essentiel dans les modèles globaux du métabolisme car elles empêchent certaines réactions de fonctionner dans un sens thermodynamiquement impossible *in vivo*. Leur prise en compte dans les modèles élimine de ce fait un grand nombre de distributions de flux impossibles du point de vue de la thermodynamique – par exemple la régénération « gratuite » d'ADP en ATP par des réactions à l'irréversibilité ignorée.

Les grandes bases de données métaboliques ne spécifiant pas de manière rigoureuse la réversibilité des réactions, il revient au modélisateur de recueillir les informations nécessaires à l'application de ces contraintes. Lors de la reconstruction du modèle d'*A. baylyi*, nous avons recherché ces informations dans la littérature et la base de données enzymatiques BRENDA. Lorsque ces informations n'étaient pas disponibles, nous nous sommes alors appuyés sur un ensemble de règles simples définies par Ma & Zeng (2003) à partir de considérations thermodynamiques. Ces règles consistent à rendre irréversibles des réactions mettant en jeu un ensemble de métabolites hautement énergétiques (Ma & Zeng 2003). Il est important de noter que, en général, la contrainte de réversibilité est appliquée au modèle quelles que soient les conditions considérées. Les irréversibilités retenues doivent donc être valables de manière très générale.

6.2.7 Associations gènes-réactions

Afin de prédire l'effet de l'inactivation d'un ou plusieurs gènes sur le fonctionnement du métabolisme, le modèle doit pouvoir tenir compte des dépendances entre les gènes et les réactions pour déterminer quelles réactions sont inactivées par la perturbation génétique. Pour ce faire, à chaque réaction est associée une règle booléenne appelée GPR (pour « Gene-Protein-Reaction associations », introduites par Reed *et al* (2003)) exprimant formellement la dépendance de la réaction aux gènes codant pour ses enzymes (voir revue sur les modèles section 3.2.1) : des gènes codant pour des sous-unités d'un complexe enzymatique sont liés par une règle ET (ils sont tous requis), tandis que des gènes codant pour des enzymes alternatives sont liés par la règle OU (l'un ou l'autre est requis). À titre d'exemple, la réaction de synthèse du glutamate à partir de glutamine et d'α-ketoglutarate (glutamate synthase) est catalysée chez *A. baylyi* par deux complexes enzymatiques distincts ; sa GPR dans le modèle est donnée par la formule :

(ACIAD3349 and ACIAD3350) or (ACIAD2525 and ACIAD2526 and ACIAD2527)

La construction des GPR à grande échelle est rendue difficile par la nécessité de déterminer les complexes enzymatiques. Même si la participation du produit d'un gène à un complexe plus grand est parfois mentionnée textuellement dans les annotations (« enzyme subunit »), l'information n'est en général pas suffisamment explicite et organisée pour inférer automatiquement tous les complexes enzymatiques. Il est donc nécessaire de s'appuyer sur la connaissance des complexes identifiés dans la bactérie (via la littérature correspondante) ou les bactéries proches (par homologie).

Certaines ressources répertorient de manière organisée les complexes identifiés dans les organismes modèles, notamment EcoCyc pour *E. coli* (Keseler et al. 2009). Afin d'accélérer la construction des GPR pour *A. baylyi*, nous avons cherché à exploiter automatiquement cette connaissance en développant un programme reconstituant les complexes d'*A. baylyi* par homologie à ceux d'*E. coli*. Pour chaque complexe d'*E. coli*, ce programme recherche dans *A. baylyi* des homologues

proches[49] pour chacun des gènes codant pour les protéines du complexe. Si un homologue pour chacun des gènes est retrouvé dans *A. baylyi*, le complexe est alors reconstitué avec ces homologues en conservant la même structure (voir Figure 30). Nous avons implémenté ce programme en utilisant la librairie CYCLONE développée par d'autres membres de notre groupe (Le Fèvre et al. 2007) pour interroger EcoCyc et créer les complexes dans AcinetoCyc.

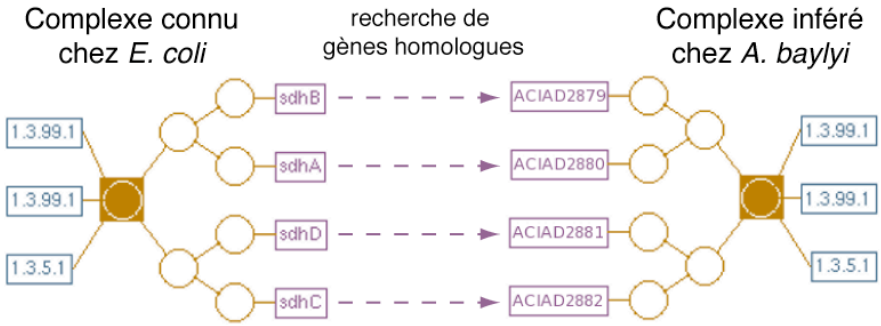


Figure 30. Inférence automatique de complexes par homologie aux complexes de *E. coli*. Pour chaque complexe de *E. coli*, un homologue[49] à chacun des gènes impliqués est recherché chez *A. baylyi*. Si tous les gènes sont retrouvés, un complexe de même structure est alors inféré chez *A. baylyi*. La représentation sous forme de graphe est issue de AcinetoCyc.

L'exécution du programme a permis d'inférer la présence de 310 complexes chez *A. baylyi*, pour 821 complexes examinés chez *E. coli* (version 9 d'Ecocyc, de 2005). Parmi les complexes inférés, 241 sont homomériques et 69 hétéromériques, ces derniers impliquant donc plusieurs gènes et pouvant donner lieu à des règles ET dans les GPR.

L'inférence des complexes étant basée uniquement sur des critères d'homologie et ne tenant pas compte des annotations fonctionnelles, nous avons (1) vérifié que les complexes homologues catalysaient bien les mêmes réactions dans EcoCyc et AcinetoCyc et (2) corroboré l'existence du complexe avec l'annotation des gènes impliqués. Seuls les complexes vérifiés ont finalement été intégrés dans la dernière version d'AcinetoCyc. Dans un deuxième temps, nous avons poursuivi ce travail de curation manuelle en cherchant à identifier les complexes enzymatiques non prédits (car n'existant pas chez *E. coli*). Nous avons pour cela exploré les annotations

[49] Nous avons utilisé comme critères d'homologie un seuil minimal de similarité de 45%, une longueur d'alignement d'au moins 80% de la séquence la plus courte et une contrainte de meilleur alignement réciproque (« Bidirectional Best Hit »).

textuelles (en recherchant les annotations possédant les mots clés « subunit », « complex », ou « component ») et la littérature relative aux voies métaboliques étudiées chez *A. baylyi*. Une grande proportion des complexes non inférés était constituée de transporteurs absents chez *E. coli* (majoritairement des transporteurs ABC), illustrant la différence de ressources nutritionnelles utilisées par *A. baylyi* et *E. coli*. Pour faciliter spécifiquement la reconstruction de ces complexes, nous avons utilisé la ressource TransportDB qui décrit explicitement les sous-unités de chaque transporteur et prédit leur présence dans le génome (Ren et al. 2004).

Le modèle d'*A. baylyi* comprend au total 63 complexes distincts, dont 28 furent inférés directement par homologie à *E. coli*. Parmi les complexes déterminés « manuellement », 26 possèdent une fonction enzymatique et les 9 restant assurent des fonctions de transport.

6.2.8 Composition de la biomasse

L'introduction d'une réaction de biomasse dans les modèles à base de contraintes permet de modéliser les effets de la croissance sur les flux métaboliques. Pour cela, cette réaction consomme les métabolites précurseurs de la biomasse, simulant quantitativement leur utilisation par la cellule pour construire les macromolécules nécessaires à son fonctionnement et à sa croissance. Grâce à la contrainte de conservation de la masse, les voies de biosynthèses de ces précurseurs sont alors contraintes de fonctionner avec un flux permettant de répondre à cette consommation (voir revue sur les modèles, section 3.2.1).

La définition de la réaction de biomasse se heurte cependant à une difficulté majeure : la composition de la biomasse et les consommations en précurseurs dépendent sensiblement des conditions de croissance, notamment l'environnement extérieur, la vitesse de croissance atteinte ou les éventuelles perturbations génétiques. À chaque condition de croissance devrait ainsi correspondre une réaction de biomasse d'équation bilan particulière, rendant a priori difficile leur utilisation pour effectuer des prédictions de croissance dans un grand ensemble de conditions. Dans un souci de simplification, deux grandes catégories de réactions de biomasse sont utilisées, en fonction des types de prédictions recherchées. D'une part, une première catégorie de

réactions de biomasse est formulée pour prédire quantitativement la croissance de l'organisme sauvage[50] dans des conditions expérimentales classiques. La consommation des précurseurs est alors typiquement estimée à partir d'études de la composition de la biomasse pour ces conditions expérimentales : on suppose que la consommation des précurseurs est uniquement due à la dilution provoquée par la croissance[51] et est donc égale à la quantité de précurseurs employés dans la cellule multipliée par le taux de croissance. D'autre part, la seconde catégorie de réactions de biomasse est utilisée pour prédire qualitativement l'aptitude à croître de mutants de délétion. Seuls sont alors retenus dans la réaction de biomasse les précurseurs essentiels à la survie de la cellule. Ces réactions de biomasse sont généralement formulées en soustrayant des précurseurs non vitaux aux réactions de biomasse de la première catégorie.

Dans le cas d'*A. baylyi*, nous avons défini pour la souche sauvage une réaction de biomasse « quantitative » que nous avons ensuite réduite pour prédire les phénotypes de croissance des mutants (avec quelques améliorations opérées par la suite grâce aux comparaisons avec les phénotypes expérimentaux, voir partie suivante). Étant donné l'absence de données spécifiques à *A. baylyi*, nous nous sommes basés sur des études de la composition de la biomasse de plusieurs souches du genre *Acinetobacter* en supposant les résultats extrapolables à *A. baylyi.*

Afin de simplifier la « gestion » de la réaction de biomasse, nous l'avons décomposée en plusieurs sous-réactions. Tout d'abord un ensemble de réactions assemblant chacune un type de macromolécules à partir de ses précurseurs, puis une réaction globale consommant toutes les macromolécules pour former la biomasse « totale ». Nous avons retenu les macromolécules suivantes pour *A. baylyi* : protéines, ADN, ARN, acides gras libres, triglycérides, wax-esters, phospholipides, lipopolysaccharides, polysaccharides libres, peptidoglycanes et cofacteurs (ces derniers ne constituant pas une macromolécule, mais rassemblant sous la forme d'un métabolite virtuel un ensemble de cofacteurs dont la présence est nécessaire à la survie de la cellule). Nous détaillerons dans les paragraphes ci-dessous les

[50] Non modifié génétiquement.
[51] On néglige dans ce cas les cycles potentiels de dégradation/synthèse des précurseurs de biomasse devant leur consommation par la croissance.

compositions des macromolécules en précurseurs ainsi que la composition globale en macromolécules retenues pour le modèle d'*A. baylyi*. Un compte rendu exhaustif et détaillé des résultats extraits de la littérature est disponible sous la forme d'un tableur Excel à l'adresse : http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S4.xls .

Composition globale

La composition de la biomasse en macromolécules est, dans la plupart des études expérimentales, donnée en pourcentage de la masse sèche totale de la biomasse. Le modèle exprimant les quantités de métabolites converties en millimoles (l'unité de flux choisie étant le $mmol.h^{-1}.(g\ DW)^{-1}$, voir 1.2.4) et non pas en masse, une étape de conversion de grandeur est nécessaire pour intégrer ces résultats expérimentaux. Pour ce faire, nous avons ajouté au modèle pour chaque métabolite dont l'expression en masse est nécessaire (notamment les macromolécules), une réaction créant un métabolite exprimé en masse :

$$\text{métaboliteA} \leftrightarrow MM_A\ \text{métaboliteA}_W$$
$$(mmol) \quad \leftrightarrow (g.mmol^{-1}) \quad (g)$$

où MM_A est la masse molaire du métabolite et l'indice $_W$ indique le métabolite exprimé en masse. Ce dernier est inclu uniquement pour modéliser la formation de biomasse et ne participe à aucune réaction biochimique.

La réaction globale de biomasse peut de ce fait être directement exprimée à partir des proportions massiques en macromolécules :

$$p_A\ \text{macromoléculeA}_W + p_B\ \text{macromoléculeB}_W + \ldots \rightarrow 1\ \text{biomasse}_W$$

où les p_A, p_B, etc. sont les proportions massiques de chacune des macromolécules. Il est important de noter que le flux de cette réaction s'exprime alors avec l'unité $(g\ DW).h^{-1}.(g\ DW)^{-1} = h^{-1}$, représentant ainsi directement le taux de croissance de l'organisme.

Nous sommes appuyés sur diverses études présentées dans la littérature pour déterminer la composition globale en macromolécules d'*A. baylyi* (voir Tableau 5). Abbott *et al* (1974) et du Preez *et al* (1984) ont mesurés la composition de la

biomasse en protéines, ADN, ARN et carbohydrates pour *A. calcoaceticus* cultivée sur des milieux minimaux d'éthanol et d'acetate. Nous avons retenu une moyenne de leurs mesures pour un taux de croissance de 0,6 pour déterminer les proportions massiques de protéines, ADN et ARN. Makula *et al* (1975) ont déterminé la composition en lipides de la bactérie *Acinetobacter* sp. HO1-N cultivée sur milieu riche et hexadecane. Nous avons retenu la composition mesurée sur milieu riche, le milieu hexadecane provoquant une accumulation de lipides spécifique à ce type de milieu. Les milieux étudiés par la suite se rapprochent en effet plus du milieu riche du point de vue des lipides. Thorne *et al* (1973) et Scott *et al* (1976) ont étudié la composition des membranes (interne pour les deux articles et externe pour Thorne *et al* (1973)) de deux bactéries du genre *Acinetobacter* cultivées sur milieux riches. Nous avons retenus leurs résultats pour déterminer les compositions en polysaccharides, lipopolysaccharides et peptidoglycane. Enfin, nous avons estimé la masse de cofacteurs et petites molécules par analogie à *E. coli* (Neidhardt & Umbarger 1996).

Macromolécule	Proportion massique	Référence
Peptidoglycane	2,8%	(Thorne et al. 1973)
Polysaccharides libres	4,1%	(Abbott et al. 1974; du Preez et al. 1984; Thorne et al. 1973)
Lipopolysaccharides	0,3%	(Scott et al. 1976; Thorne et al. 1973)
Phospholipides	4,1%	(Makula et al. 1975)
Wax esters	0,6%	(Makula et al. 1975)
Triglycérides	0,2%	(Makula et al. 1975)
Acides gras libres	0,2%	(Makula et al. 1975)
ADN	3,2%	(Abbott et al. 1974; du Preez et al. 1984)
ARN	20%	(Abbott et al. 1974; du Preez et al. 1984)
Protéines	63,3%	(Abbott et al. 1974; du Preez et al. 1984)
Cofacteurs	3,2%	(Neidhardt & Umbarger 1996)

Tableau 5. Composition massique de la biomasse en macromolécules retenue pour le modèle d'*A. baylyi*.

La consommation d'énergie requise par la croissance de la cellule est prise en compte par deux termes énergétiques d'hydrolyse d'ATP : un terme inclu dans la réaction de biomasse (terme proportionnel au taux de croissance) et un terme de flux fixe (terme indépendant du taux de croissance) (voir revue sur les modèles, section 3.2.1). Ne disposant pas de mesures expérimentales de rendements de croissance suffisamment précises pour *A. baylyi*, nous avons adopté celles du modèle d'*E. coli* (Reed et al. 2003) :

- terme associé à la croissance : 40 mmol.(g DW)$^{-1}$ d'ATP, inclus dans la réaction globale de biomasse
- terme indépendant de la croissance : 10 mmol.h^{-1}.(g DW)$^{-1}$ d'ATP, modélisé sous la forme d'une réaction d'hydrolyse d'ATP de flux fixe.

Nous verrons plus loin que les prédictions de taux de croissance sont d'ailleurs peu sensibles à ces paramètres (voir 7.2).

Protéines

Nous n'avons trouvé aucune étude de la composition en acides aminés de bactéries du genre *A. baylyi*. Nous nous sommes alors reportés sur la composition d'*E. coli* pour construire la réaction d'assemblage de la macromolécule protéine à partir des acides aminés (Neidhardt & Umbarger 1996) (voir Tableau 6). Nous avons tenu compte directement, dans cette réaction, du coût énergétique de traduction (polymérisation et correction d'erreur) à hauteur de 4,2 ATP hydrolysé par acide aminé (Oliveira et al. 2005).

Acide aminé	Composition molaire
alanine	13%
arginine	3%
asparagine	4%
aspartate	4%
cystéine	2%
glutamate	4%
glutamine	4%
glycine	20%
histidine	1%
isoleucine	5%
leucine	7%
lysine	5%
méthionine	2%
phénylalanine	2%
proline	4%
sérine	4%
thréonine	5%
tryptophane	1%
tyrosine	2%
valine	8%

Tableau 6. Composition moyenne des protéines en acides aminés retenue pour le modèle d'*A. baylyi.*

Acides nucléiques

Nous avons déterminé la composition en bases nucléotidiques des macromolécules ADN et ARN à partir du pourcentage GC pour l'ADN (voir Tableau

7) et d'un décompte moyen des bases contenues dans les ARNr, ARNt et ARNm (voir Tableau 8). ADN et ARN sont assemblés à partir de nucléotides triphosphates (dNTP et NTP) et des coûts énergétiques supplémentaires (d'assemblage et correction d'erreur) de respectivement 1,37 et 0,4 ATP hydrolysés par nucléotide sont ajoutés aux réactions assemblant ces macromolécules (Oliveira et al. 2005).

Nucléotide	Proportion molaire
dAMP	30%
dTMP	30%
dGMP	20%
dCMP	20%

Tableau 7. Composition moyenne en nucléotide de l'ADN retenue pour le modèle d'*A. baylyi.*

Nucléotide	Proportion molaire
AMP	22%
UMP	26%
GMP	22%
CMP	30%

Tableau 8. Composition moyenne en nucléotide de l'ARN retenue pour le modèle d'*A. baylyi.*

Lipides

Nous avons exploité trois études de la composition des lipides de bactéries du genre *Acinetobacter* pour déterminer une composition moyenne en termes de chaînes carbonées dans les différents lipides retenus (Makula et al. 1975; Scott et al. 1976; Thorne et al. 1973). Seules des chaînes de longueurs paires furent détectées (en accord avec les voies biosynthétiques identifiées), avec éventuellement une insaturation (voir Tableau 9). Pour chaque type de lipide, nous avons alors constitué une macromolécule « générique » à partir des lipides de chaînes carbonées définies selon les proportions décrites dans le Tableau 9.

Chaîne carbonée	Phospho-lipides	Triglycéri-des	Wax esters - a. gras	Wax esters - alcool	Acides gras libres
14:0	0%	2%	0%	0%	6%
16:0	10%	72%	24%	5%	28%
16:1	20%	5%	32%	48%	57%
18:0	10%	6%	0%	0%	0%
18:1	60%	15%	44%	47%	9%

Tableau 9. Compositions molaires moyennes en types de chaînes carbonées des différents lipides pris en compte dans la biomasse du modèle d'*A. baylyi*.

L'étude de Scott et al (1976) a également permis de déterminer la composition en phospholipides de la membrane. Nous en avons repris les résultats (sur milieu riche) dans le modèle (voir Tableau 10).

Phospholipide	Proportion molaire
phosphatidyl-glycerol	10%
phosphatidyl-ethanolamine	73%
cardiolipin	17%

Tableau 10. Composition moyenne des phospholipides retenue dans le modèle d'*A. baylyi*.

Paroi cellulaire

Nous avons tenu compte de deux types de macromolécules constituant la paroi cellulaire (en plus des phospholipides traités ci-dessus) : le peptidoglycane et les lipopolysaccharides (LPS).

S'agissant du peptidoglycane, nous avons retenu dans le modèle une composition typique des bactéries gram-négatives, à savoir un enchaînement de résidus N-acétyl-glucosamine et acide N-acétyl-muramique liés au pentapeptide L-ala / D-glu / diaminopimelate / D-ala / D-ala (voir Figure 31).

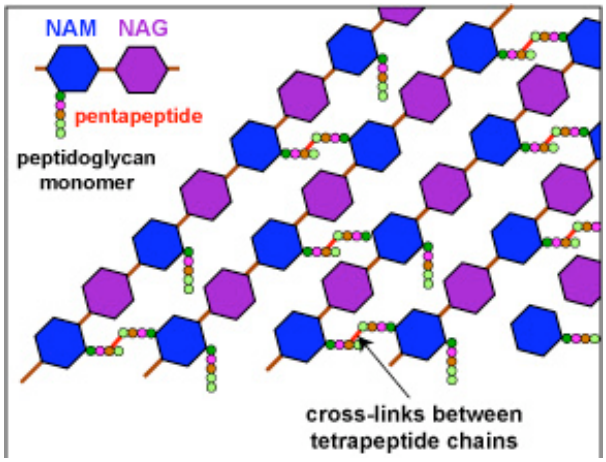


Figure 31. Agencement des chaînes de peptidoglycane dans la paroi cellulaire. NAG, N-acétyl-glucosamine ; NAM, acide N-acétyl-muramique.
Figure extraite de http://student.ccbcmd.edu/~gkaiser/

Les molécules de LPS se composent d'un lipide A et d'un cœur d'oligosaccharides. Nous avons exploité l'analyse expérimentale de la paroi cellulaire réalisée par Thorne *et al* (1973) pour définir la composition en lipides de chaque molécules de lipide A (voir Tableau 11). Nous avons supposé que deux glucosamines portaient les acides gras du lipide A.

Type d'acide gras	Nombre de molécules par lipide A	Composition	Type de chaîne carbonée
Beta-hydroxy	4	56%	12:betaOH
		44%	14:betaOH
Classique	2	62%	12:0
		18%	16:0
		10%	18:0
		10%	18:1

Tableau 11. Composition lipidique du lipide A retenue dans le modèle d'*A. baylyi.*

Les travaux de Thorne et al (1973) et Bryan et al (1986) nous ont permis de définir la composition des polysaccharides attachés au LPS. Pour chaque LPS, nous avons associé au lipide A deux molécules de KDO ainsi que 5 molécules d'hexoses, composés à parts égales de glucose, rhamnose et mannose.

Nous avons également adopté une composition équirépartie de ces trois hexoses pour les polysaccharides libres (Bryan et al. 1986).

Cofacteurs

Les cofacteurs métaboliques sont inclus dans la réaction de biomasse dans le but de tenir compte de leur essentialité pour la croissance des cellules. En effet, à l'état stationnaire, la régénération des cofacteurs permet au modèle d'utiliser ces derniers sans recourir à leur synthèse. Celle-ci est néanmoins vitale pour répondre à leur dilution par la croissance.

Nous avons déterminé la liste des cofacteurs essentiels par analogie au modèle d'*E. coli* (Reed et al. 2003) (voir Tableau 12).

coenzyme A
fad
fmn
folate (THF)
heme
nad
nadp
pyridoxal-5p
s-adenosylmethionine
siroheme
thiamin
ubiquinone-8
undecaprenyl-pp

Tableau 12. Cofacteurs essentiels pris en compte dans le modèle.

Prédiction qualitative des phénotypes de croissance de mutants

De manière à prédire qualitativement l'aptitude à croître de la bactérie ou d'un de ses mutants sur un environnement donné (voir application partie suivante), nous avons dérivé des réactions de biomasse présentées ci-dessus une liste réduite de précurseurs que nous avons considéré comme essentiels à la croissance (voir Tableau 13). Nous avons prédit les phénotypes qualitatifs de croissance en analysant ensuite la simple productibilité de chacun des métabolites de cette liste.

tetradecanoate (C14:0)	ttp
hexadecanoate (C16:0)	triacylglycerol
cis-hexadec-7-enoate (C16:1)	l-1-phosphatidyl-ethanolamine
cis-octadec-9-enoate (C18:1)	l-1-phosphatidyl-glycerol
gdp-mannose	cardiolipin
udp-d-glucose	fad
dtdp-rhamnose	fmn
kdo2-lipid a	s-adenosyl-l-methionine
peptidoglycane	thiamine-pyrophosphate
protéine	pyridoxal 5'-phosphate
ctp	nadp
gtp	ubiquinone(40)
utp	siroheme
atp	heme o
datp	nad
dctp	coenzyme A
dgtp	thf

Tableau 13. Liste des précurseurs de biomasse utilisés pour prédire la capacité à croître de l'organisme.

7 Le modèle d'*Acinetobacter baylyi*

À la suite de ce processus de reconstruction, nous avons obtenu un modèle global du métabolisme d'*A. baylyi* que nous avons appelé iAbaylyi^{v1}. L'indice ^{v1} indique que ce modèle constitue une première version reconstruite à partir de l'ensemble des connaissances initialement disponibles sur le métabolisme d'*A. baylyi*. Au fur et à mesure des améliorations apportées au modèle grâce à l'exploitation de données expérimentales additionnelles (voir partie suivante), nous distinguerons les versions du modèle en incrémentant cet indice.

Une grande partie des caractéristiques du modèle iAbaylyi^{v1} étant présentée dans l'article Durot *et al* (2008) inclus dans la partie suivante, nous avons choisi d'exposer dans ce chapitre uniquement quelques compléments utiles. Ainsi, nous donnerons tout d'abord au lecteur un aperçu plus détaillé des voies métaboliques modélisées, puis

nous effectuerons quelques analyses sur les prédictions quantitatives de croissance, enfin nous présenterons quelques détails techniques sur l'utilisation du modèle.

7.1 Composition métabolique globale

Le Tableau 14 présente l'ensemble des catégories fonctionnelles et voies métaboliques prises en compte dans le modèle iAbaylyi[v1], ainsi que les nombres de réactions et gènes impliqués dans chacune d'entre elles.

Catégorie fonctionnelle	Voie métabolique	Réactions	Gènes
TOTAL		**970**	**787**
Amino acids metabolism		**139**	**158**
	Alanine biosynthesis	3	7
	Arginine biosynthesis	10	12
	Arginine degradation	5	5
	Aspartate / asparagine biosynthesis	2	3
	Aspartate / asparagine degradation	2	3
	Betaine biosynthesis	2	2
	Biomass assembly	1	0
	Chorismate biosynthesis	7	8
	Cysteine biosynthesis	3	6
	Glutamate / glutamin biosynthesis	4	11
	Glutamate degradation	2	4
	Glycine biosynthesis	1	1
	Histidine metabolism	11	10
	Isoleucine biosynthesis	5	8
	Leucine and valine biosynthesis	8	12
	Lysine biosynthesis	7	7
	Methionine metabolism	8	8
	Ornithine degradation	1	1
	Other	3	3
	Phenylalanine / tyrosine biosynthesis	7	6
	Proline biosynthesis	4	3
	Proline degradation	2	1
	Serine / glycine biosynthesis	4	4
	Sulfur metabolism	3	5
	Threonine biosynthesis	5	7
	tRNA charging pathway	22	26
	Tryptophan biosynthesis	5	7
	Urea degradation	2	4
Central metabolism		**70**	**109**
	Acetate metabolism	3	6
	Biomass assembly	1	0
	Citrate cycle (TCA)	10	19
	Entner-Doudoroff pathway	6	6
	Glycerol metabolism	3	3
	Glycolysis/gluconeogenesis	12	14
	Glyoxylate shunt	2	2
	Maintenance flux	1	0
	Malonate metabolism	1	7
	Other	1	0
	Pentose phosphate	5	4
	Propionate metabolism / methylcitrate pathway	7	6

	Pyruvate metabolism	5	7
	Radicals detoxification	3	6
	Respiration	10	35
Cofactors biosynthesis		**107**	**103**
	Biomass assembly	1	0
	Biotin biosynthesis	5	5
	Coenzyme A biosynthesis	9	8
	FMN / FAD biosynthesis	9	7
	Folate metabolism	16	16
	Glutathione biosynthesis	2	2
	Heme / siroheme biosynthesis	14	12
	NAD / NADP biosynthesis	11	10
	Other	4	8
	Polyisoprenoids biosynthesis	14	13
	Pyridoxal 5P biosynthesis	7	7
	Thiamin biosynthesis	6	9
	Ubiquinone biosynthesis	9	7
Degradation pathways		**181**	**163**
	3,4-dihydroxyphenylpropionate degradation	2	2
	3-hydroxy-isobutyrate degradation	2	5
	Acetoacetate metabolism	4	6
	Aldoxime / nitrile degradation	3	4
	Alkane degradation	4	5
	Allantoate degradation	2	2
	Anthranilate degradation	1	3
	Benzyl compounds degradation	7	9
	Beta-ketoadipate degradation	2	9
	Butanediol and acetoin degradation	3	6
	Butyric acid metabolism	2	12
	Caffeate degradation	4	3
	Catechol degradation	3	3
	Chlorogenate / quinate degradation	4	5
	Coumarate degradation	5	5
	Dicarboxylates degradation	24	7
	Ethanol metabolism	2	7
	Fatty acids degradation	58	43
	Ferulate / vanillate degradation	6	6
	Fructose utilisation	1	1
	Galactarate / glucarate degradation	4	4
	Glyceraldehyde degradation	3	5
	Glycolaldehyde degradation	1	2
	Lactate utilisation	3	2
	Methylglyoxal degradation	2	2
	Nitrogen assimilation	3	4
	Other	6	8
	Protocatechuate degradation	4	8
	Ribose utilisation	1	1
	Salicyl compounds degradation	5	8
	Sarcosine degradation	1	4
	Shikimate degradation	2	2
	Sulfonate degradation	4	11
	UDP-glucose utilisation	3	5
Lipid, cell wall biosynthesis		**141**	**76**
	Biomass assembly	4	0
	Fatty acids biosynthesis	64	20
	KDO-lipid A biosynthesis	16	13

	Lipoate biosynthesis	1	2
	Peptidoglycan biosynthesis	14	14
	Phospholipids biosynthesis	9	9
	Polysaccharides biosynthesis	11	11
	TAG metabolism	10	5
	Wax ester metabolism	12	4
Nucleotide metabolism		**88**	**64**
	Biomass assembly	2	0
	Other	1	3
	Purine biosynthesis - de novo	27	24
	Purine biosynthesis - salvage pathways	22	12
	Pyrimidine biosynthesis - de novo	22	28
	Pyrimidine biosynthesis - salvage pathways	14	8
Transport		**133**	**127**
Exchange fluxes		**109**	**0**
Unbalance fluxes[52]		**2**	**0**

Tableau 14. Répartition du nombre de gènes et de réactions dans le modèle iAbaylyi^{v1} selon les catégories fonctionnelles et les voies métaboliques. Chaque réaction est assignée à une unique voie métabolique. Certains gènes sont associés à plusieurs réactions ; ils peuvent de ce fait être comptés dans plusieurs catégories ou voies.

Le modèle compte un total de 970 réactions, comprenant 111 réactions d'échanges (catégories « Exchange fluxes » et « Unbalance fluxes »), 9 réactions de biomasse (voies métaboliques « Biomass assembly »), 133 transporteurs et 717 réactions purement métaboliques. Ces dernières se répartissent dans les six grandes catégories fonctionnelles du métabolisme que nous avons définies lors de la reconstruction, illustrant le fait que le modèle prend en compte le métabolisme dans sa globalité. Il comprend aussi bien les voies dédiées à l'anabolisme des constituants de la cellule (catégories « Amino acids metabolism », « Cofactors biosynthesis », « Lipid, cell wall biosynthesis » et « Nucleotide metabolism ») que celles impliquées dans le catabolisme des nutriments et la génération d'énergie (catégories « Central metabolism » et « Degradation pathways » principalement).

Les particularités métaboliques d'*A. baylyi* se retrouvent dans le contenu en voies métaboliques du modèle (voir Tableau 14). Ses remarquables capacités cataboliques sont en effet reflétées par le grand nombre de réactions et la large diversité de voies regroupées dans la catégorie « Degradation pathways ». Ses capacités à synthétiser et

[52] Les réactions de la catégorie « Unbalance fluxes » sont des réactions d'échanges de métabolites intracellulaires ayant été introduites pour relaxer la contrainte de conservation de la masse. Elles concernent en l'occurrence le proton (voir 6.2.3) et le cofacteur s-adenosyl-4-methylthio-2-oxobutanoate (voir 6.2.1), tous deux localisés dans le cytoplasme.

dégrader de nombreux lipides se reflètent également dans le modèle : un nombre significatif de réactions y sont consacrés et une grande variété de lipides peut être métabolisée (acides gras avec ou sans insaturation, acides dicarboxyliques, wax esters, triglycérides notamment).

Comme évoqué plus haut à propos du processus de reconstruction (voir 6.2.5), la création de réactions spécifiques pour les enzymes à larges spectres de substrats augmente sensiblement le nombre de réactions dans le modèle. En témoigne la forte représentation des voies métaboliques « Fatty acids biosynthesis », « Fatty acids degradation » et « Dicarboxylates degradation » dans le modèle (146 réactions à elles seules). Ces voies métaboliques contiennent des réactions ayant été générées à partir de réactions génériques selon le processus décrit plus haut (voir 6.2.5). Ces réactions étant catalysées par les mêmes enzymes, le nombre de gènes impliqués dans ces voies reste donc limité ; la voie « Fatty acids degradation » déroge cependant à cette règle (43 gènes inclus) du fait du très grand nombre d'isozymes identifiées pour catalyser certaines des réactions de cette voie (nombreuses acyl-coa dehydrogenase, enoyl-coa hydratase et 3-oxoacyl-coa thiolase notamment).

7.2 Prédictions quantitatives de croissance

Durant les travaux de notre thèse, nous nous sommes majoritairement concentrés sur les prédictions qualitatives de phénotypes de croissance (prédiction de la simple aptitude à croître, développements et résultats présentés dans les parties suivantes). Les modèles à base de contraintes permettent cependant d'effectuer des prédictions quantitatives de taux de croissance, prédictions utilisées avec intérêt pour un nombre significatif d'autres applications (voir 3.2.1). Afin d'évaluer la capacité de notre modèle à réaliser correctement ce type de prédictions, nous avons effectué deux études relativement simples sur les taux de croissance prédits.

7.2.1 Comparaison des prédictions de taux de croissance à des mesures expérimentales

Nous avons tout d'abord comparé les prédictions de taux de croissance du modèle iAbaylyi^{v1} à des mesures expérimentales. L'équipe Thesaurus du Genoscope a pour cela réalisé une culture suivie d'*A. baylyi* en milieu minimal liquide contenant du

glutamate comme seule source de carbone[53]. À chaque point de temps (toutes les heures) fut effectué un prélèvement à partir duquel la densité optique à 600 nm (DO) et la concentration en glutamate[54] furent déterminées (voir Figure 32).

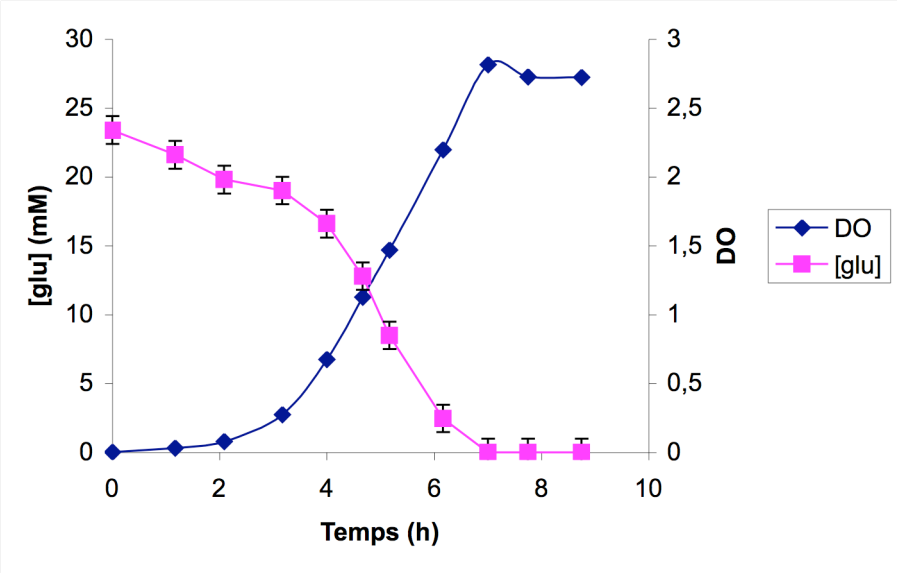


Figure 32. Évolution de la densité optique (DO) et de la concentration en glutamate dans le milieu ([glu]) au cours de la croissance suivie d'*A. baylyi*.

En utilisant une relation entre DO et masse sèche de cellules déterminée auparavant dans les mêmes conditions expérimentales, nous avons calculé les flux moyens de consommation de glutamate (par unité de masse sèche de cellules) et les taux de croissance moyens à chaque point de temps. Ces résultats sont présentés dans la Figure 33. Le faible nombre de points de temps[55] ne nous a donné accès qu'à des moyennes relativement larges pour ces grandeurs, entachées pour les premiers points de temps de marges d'erreurs importantes[56] (voir Figure 33). Cependant, aux points de temps correspondant à des périodes de croissance stable (notamment en fin de phase exponentielle, points 3h, 4h et 5h), les valeurs calculées de flux de glutamate et de taux de croissance peuvent être considérées comme suffisamment fiables pour être comparées au moins grossièrement aux prédictions du modèle.

[53] Milieu minimal contenant de l'ammoniac (NH_3) comme source d'azote.
[54] La concentration de glutamate fut déterminée par test enzymatique en le faisant réagir avec le NAD à l'aide de la glutamate dehydrogenase. La quantité de NADH produite fut déterminée par mesure d'absorption à 340 nm.
[55] Exigé par les conditions de l'expérience.
[56] Dues à la normalisation par la masse mesurée de cellules, faible au début de l'expérience.

Nous avons ensuite pour chaque point de temps contraint le modèle iAbaylyi^{v1} à importer le glutamate avec le flux mesuré puis calculé le taux de croissance optimal en maximisant le flux de la réaction de biomasse (méthode de « Flux Balance Analysis » (FBA)). La Figure 33 présente ces prédictions en regard des taux de croissance mesurés.

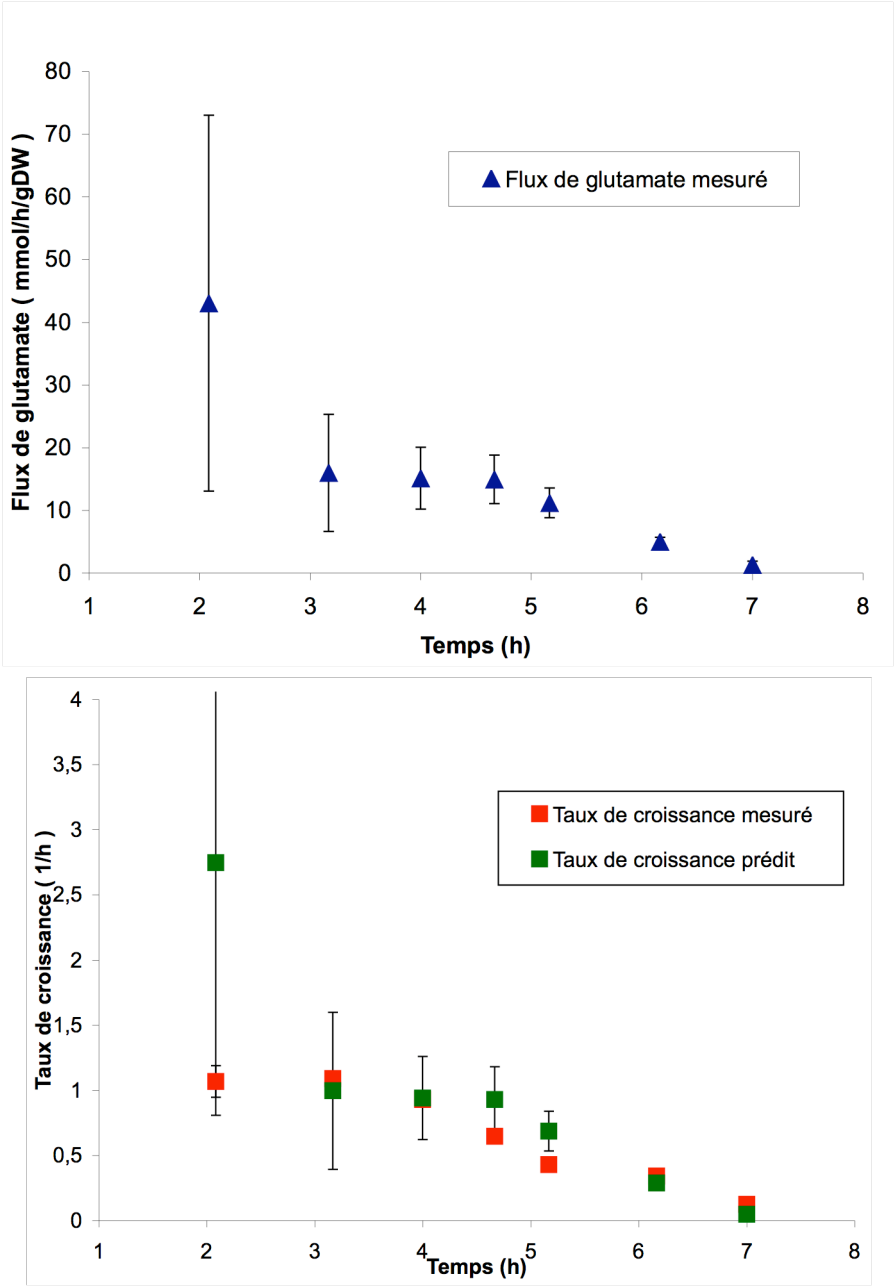


Figure 33. Flux de consommation de glutamate mesuré (en haut) et comparaison des taux de croissance mesuré et prédit (en bas). Le flux de glutamate et le taux de croissance mesuré furent déterminés à partir des mesures de concentration et de DO (la relation entre DO et masse sèche fut déterminée en parallèle, dans des conditions expérimentales similaires). Les taux de croissance furent prédits à l'aide du modèle iAbaylyi^{v1} par « Flux Balance Analysis » (Varma & Palsson 1994) en contraignant le flux de glutamate entrant aux valeurs mesurées. Les marges d'erreurs des taux de croissances prédits ont été obtenues à partir des marges d'erreurs des flux de glutamate mesurés.

Sur la période 3-5 h, le flux de glutamate consommé par cellule semble se maintenir à une valeur constante (autour de 15 mmol.h^{-1}.(g DW)$^{-1}$). Le taux de croissance prédit pour ces points de temps est par conséquent également constant (autour de 0,9 h^{-1}). Cette valeur est en bon accord avec les taux de croissance mesurés, particulièrement pour les points de temps 3 h et 4 h ; la population semble

croître de manière exponentielle à cette période. On observe ensuite une décroissance du taux de croissance mesuré, en avance par rapport à celle du taux de croissance prédit. On peut interpréter cette diminution du taux de croissance mesuré par rapport à celui prédit par le fait que la population sort probablement de son régime de croissance purement exponentiel. Il est alors probable que l'hypothèse d'exploitation optimale du métabolisme pour la croissance ne soit plus valable dans ces conditions.

Cette comparaison aux mesures expérimentales montre que les prédictions du modèle sont globalement en accord avec les taux de croissance mesurés. Cependant, une étude plus précise – impliquant un plus grand nombre de points de temps – et plus complète – utilisant différents milieux – serait nécessaire pour caractériser plus en détail les limites de validité des prédictions quantitatives du modèle (Edwards et al. 2001). Une telle étude sort en revanche du cadre des travaux de cette thèse.

7.2.2 Sensibilité des prédictions de taux de croissance aux paramètres énergétiques

Il est important de noter que les prédictions ci-dessus ont été réalisées en conservant les paramètres énergétiques (et la composition de la réaction de biomasse) déterminés initialement. Ces paramètres – le flux de maintenance associé à la croissance (« growth associated maintenance », GAM) et le flux de maintenance indépendant de la croissance (« non growth associated maintenance », NGAM) – sont typiquement déterminés à partir d'expériences de culture suivie, pour lesquelles plusieurs mesures de taux de croissance pour différents flux de consommation de nutriments distincts sont réalisées (voir 3.2.1). Ne disposant pas de suffisant de données expérimentales pour déterminer plus précisément ces paramètres, nous avons adopté ceux déterminés pour le modèle d'*E. coli* (Reed et al. 2003),

Toutefois, afin d'évaluer la sensibilité des prédictions de taux de croissance à ces paramètres, nous avons prédit par FBA les taux de croissance pour des valeurs de GAM et de NGAM variant de +/- 100% autour de leurs valeurs initiales (respectivement 40 mmol ATP/gDW and 10 mmol ATP/h/gDW). Nous avons choisi un milieu minimal supplémenté de succinate comme environnement de croissance, en variant son flux maximal d'import entre 0 et 20 mmol/h/gDW. Les deux figures suivantes présentent les résultats obtenus.

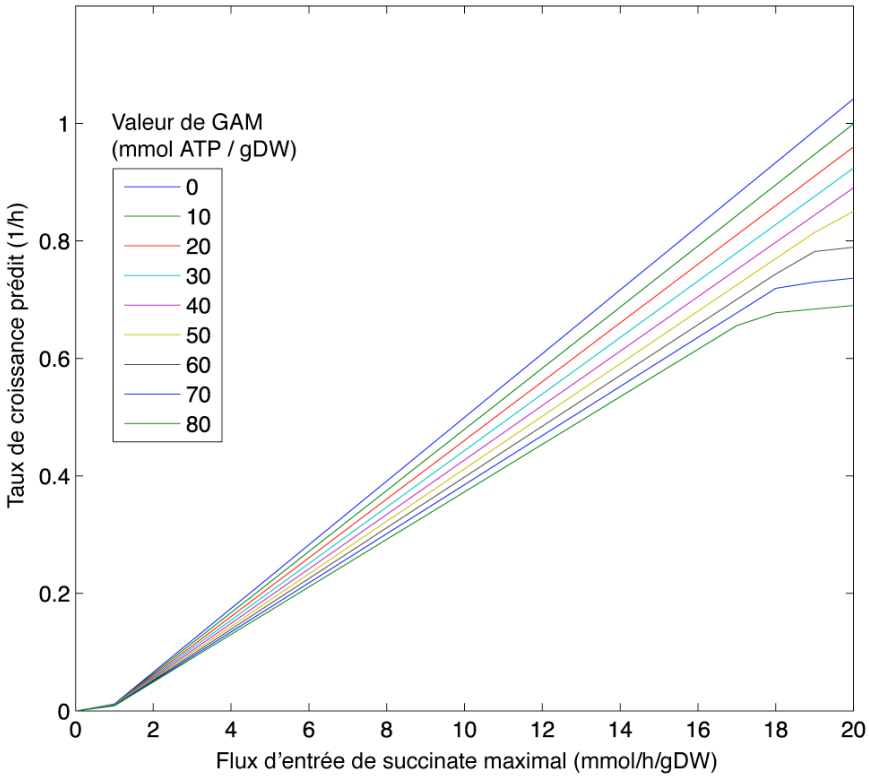


Figure 34. Effet de la variation du paramètre GAM sur les prédictions de taux de croissance. Le paramètre NGAM est fixé à 10 mmol ATP/h/gDW.

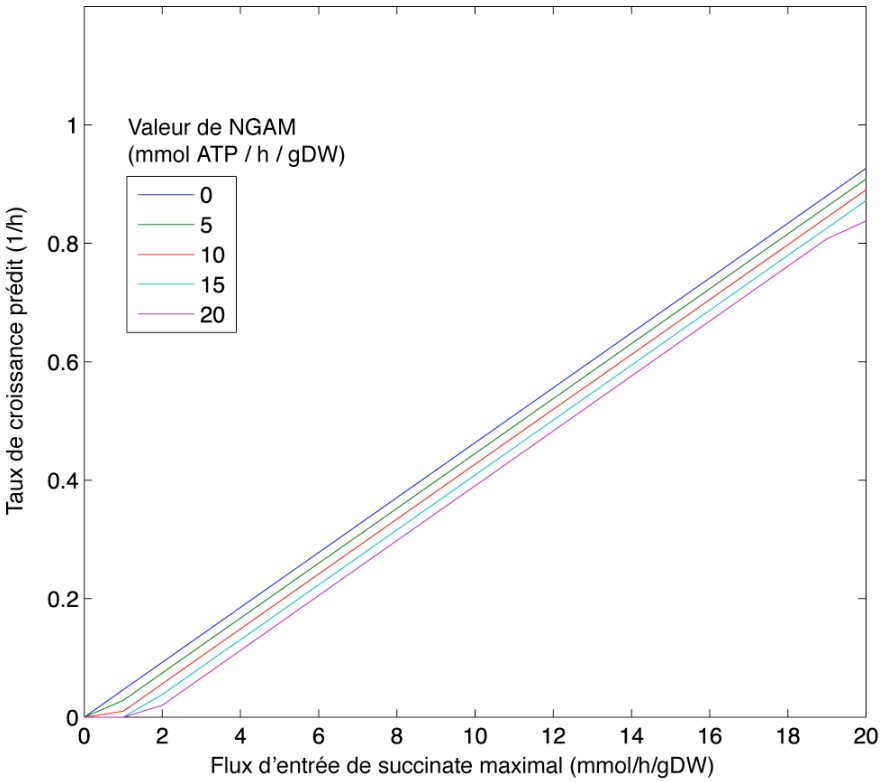


Figure 35. Effet de la variation du paramètre NGAM sur les prédictions de taux de croissance. Le paramètre GAM est fixé à 40 mmol ATP/gDW.

La variation de +/- 100% du flux de maintenance associé à la croissance (GAM) modifie le taux de croissance prédit de +/- 10%, quel que soit le flux d'entrée de succinate. Étant donné l'importance de la variation imposée au paramètre, la sensibilité des prédictions à ce paramètre est relativement réduite.

La variation du flux de maintenance non associé à la croissance (NGAM) provoque un décalage constant de +/- 0.05 h^{-1} du taux de croissance prédit, quel que soit le flux d'entrée de succinate. Bien que peu significatif pour les valeurs élevées de taux de croissance, ce décalage peut provoquer d'importantes variations pour les prédictions de taux de croissance faibles. Les prédictions réalisées avec le modèle actuel dans ces derniers cas sont donc sujettes à des incertitudes plus élevées, à la fois du fait de la forte sensibilité au paramètre NGAM mais également du fait que la validité de l'hypothèse d'optimalité du fonctionnement du métabolisme pour des taux de croissance aussi faible soit largement discutable.

7.3 Disponibilité du modèle

Pour faciliter l'utilisation du modèle, nous l'avons rendu disponible sous plusieurs formats.

Le format le plus complet est un tableur Excel contenant toutes les informations utilisées dans le modèle[57]. L'ensemble des réactions y est classé par catégories fonctionnelles et voies métaboliques, et, pour chacune d'entre elles, équation bilan, association gène-réaction (GPR), numéro EC, références bibliographiques et commentaires sont donnés. De manière à favoriser l'intégration du modèle avec d'autres ressources sur le métabolisme, le tableur fournit également les identifiants des réactions et métabolites du modèle dans les bases de données KEGG, MetaCyc et BiGG[58]. La simplicité de ce format permet de rendre compte aisément de l'information contenue dans le modèle, il nécessite cependant un travail de reformatage pour exploiter le modèle à l'aide des outils classiques de modélisation.

Pour cela, nous avons également mis à disposition le modèle au format SBML[59] (Systems Biology Markup Language) (Hucka et al. 2004). Ce format XML est exploité par un grand nombre d'outils de modélisation du métabolisme, incluant notamment YANA (Schwarz et al. 2007), CellNetAnalyzer (Klamt et al. 2007) et la COBRA Toolbox (Becker et al. 2007), qui sont spécifiquement dédiés aux modèles à base de contraintes. Le format SBML original ne spécifiant pas comment représenter les liens gène-réaction, nous avons adopté les conventions utilisées par la COBRA Toolbox et la base de données BiGG pour les inclure dans le fichier SBML[60]. Nous avons également soumis le modèle au format SBML à la base de données BioModels (Le Novère et al. 2006) ; il y est stocké sous l'identifiant MODEL1949107276.

Enfin, dans le cadre de développements informatiques menés par d'autres membres de mon groupe, nous avons inclus le modèle d'*A. baylyi* dans une interface web permettant de réaliser en ligne des prédictions de phénotypes de croissances de mutants[61]. Cette interface est étroitement associée à la base de données métabolique

[57] Téléchargeable à l'adresse : http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-s2.xls

[58] Base de données de modèles métaboliques à grande échelle : http://bigg.ucsd.edu .

[59] Téléchargeable à l'adresse : http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-s5.xml

[60] Dans chaque objet *reaction*, le lien gène-réaction est ajouté dans une *notes* de la forme :

```
<notes>
    <html:p>GENE_ASSOCIATION: aciad2449 and aciad2450 and aciad2451</html:p>
</notes>
```

[61] Disponible à l'adresse http://www.genoscope.cns.fr/nemostudio-platform/

AcinetoCyc pour visualiser les prédictions directement sur les cartes de voies métaboliques. Nous évoquerons plus en détail cet outil dans la partie suivante.

EXPLOITATION DES PHENOTYPES DE CROISSANCE DE MUTANTS PAR LE MODELE

Nous aborderons dans cette deuxième partie sur nos travaux les résultats de la confrontation du modèle d'*A. baylyi* aux phénotypes de croissance de mutants. Nous nous sommes pour cela reposés sur les ressources expérimentales de l'équipe Thesaurus présentée en introduction et avons utilisé leurs résultats de phénotypes de croissance de mutants d'*A. baylyi* pour 9 environnements minimaux distincts. De nombreux facteurs influencent les prédictions de phénotypes de croissance de mutants et peuvent être la cause d'incohérence avec les phénotypes observés. Pour faciliter l'analyse de ces incohérences, nous introduirons dans cette partie un cadre d'interprétation et montrerons que celui-ci permet de distinguer les sources d'erreurs et de guider des corrections au modèle.

Nous avons organisé cette partie en trois chapitres. Le premier reprend un article que nous avons publié dans *BMC Systems Biology* et expose l'ensemble des corrections et interprétations réalisées sur le modèle *d'A. baylyi* à partir des phénotypes expérimentaux. Il présente notamment les différentes versions du modèle progressivement obtenues au cours des étapes de correction. Cet article complète également la présentation du modèle initiée dans la partie précédente et présente une interface Web de prédictions de phénotypes pour *A. baylyi*. Dans le deuxième chapitre, nous effectuerons une synthèse des types d'interprétations réalisées à partir des incohérences de phénotypes. Enfin, nous présenterons brièvement dans le troisième chapitre une évolution récente de l'interface Web de prédiction de phénotypes prenant en compte un plus grand nombre d'organismes.

8 Article : « Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data »

BMC Systems Biology

Research article

Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data

Maxime Durot, François Le Fèvre, Véronique de Berardinis, Annett Kreimeyer, David Vallenet, Cyril Combe, Serge Smidtas, Marcel Salanoubat, Jean Weissenbach and Vincent Schachter*

Address: Genoscope (Commissariat à l'Energie Atomique) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, CP5706, 91057 Evry, Cedex, France

Email: Maxime Durot - mdurot@genoscope.cns.fr; François Le Fèvre - flefevre@genoscope.cns.fr; Véronique de Berardinis - vberard@genoscope.cns.fr; Annett Kreimeyer - akreimey@genoscope.cns.fr; David Vallenet - vallenet@genoscope.cns.fr; Cyril Combe - ccombe@genoscope.cns.fr; Serge Smidtas - smidtas@genoscope.cns.fr; Marcel Salanoubat - salanou@genoscope.cns.fr; Jean Weissenbach - jsbach@genoscope.cns.fr; Vincent Schachter* - vs@genoscope.cns.fr

* Corresponding author

Published: 7 October 2008

BMC Systems Biology 2008, **2**:85 doi:10.1186/1752-0509-2-85

Received: 23 April 2008
Accepted: 7 October 2008

This article is available from: http://www.biomedcentral.com/1752-0509/2/85

Abstract

Background: Genome-scale metabolic models are powerful tools to study global properties of metabolic networks. They provide a way to integrate various types of biological information in a single framework, providing a structured representation of available knowledge on the metabolism of the respective species.

Results: We reconstructed a constraint-based metabolic model of *Acinetobacter baylyi* ADP1, a soil bacterium of interest for environmental and biotechnological applications with large-spectrum biodegradation capabilities. Following initial reconstruction from genome annotation and the literature, we iteratively refined the model by comparing its predictions with the results of large-scale experiments: (1) high-throughput growth phenotypes of the wild-type strain on 190 distinct environments, (2) genome-wide gene essentialities from a knockout mutant library, and (3) large-scale growth phenotypes of all mutant strains on 8 minimal media. Out of 1412 predictions, 1262 were initially consistent with our experimental observations. Inconsistencies were systematically examined, leading in 65 cases to model corrections. The predictions of the final version of the model, which included three rounds of refinements, are consistent with the experimental results for (1) 91% of the wild-type growth phenotypes, (2) 94% of the gene essentiality results, and (3) 94% of the mutant growth phenotypes. To facilitate the exploitation of the metabolic model, we provide a web interface allowing online predictions and visualization of results on metabolic maps.

Conclusion: The iterative reconstruction procedure led to significant model improvements, showing that genome-wide mutant phenotypes on several media can significantly facilitate the transition from genome annotation to a high-quality model.

Background

The diversity of bacterial metabolism and the perspective of engineering applications has spurred a steep increase in both the number of sequencing projects and the volume of high throughput experiments on bacteria. The need to interpret and integrate these datasets at the systems level has triggered the development of model-based computational methods [1]. Among them, the constraint-based modeling approach (CBM) has proved to be particularly efficient at integrating large-scale *omics* datasets related to metabolism, such as growth phenotypes, metabolite concentrations, or reaction fluxes [2]. In addition to providing a structured summary of metabolism-related knowledge for a given species, a constraint-based model allows the prediction and analysis of a variety of properties resulting from topological, stoichiometric, and physiological constraints known to apply at steady-state to its global metabolic network. Applications range from studies on evolutionary or physiological properties to the design of metabolic engineering strategies for biotechnological or therapeutical purposes [3]. Nearly twenty such models have been built so far [2], typically through extensive curation work, and, for some of them, through iterative refinement processes where models were progressively improved by comparison with experimental datasets [4].

Systematic evaluation of gene essentiality has proved to be a valuable resource for investigating gene functions; knockout mutant collections have been recently built in this aim for a number of bacteria [5-8]. Rigorous analysis of their results remains a challenging task, however, as gene essentiality depends on the environmental condition and the link between genes and essential functions may be blurred by genetic or metabolic redundancy [9,10]. Genome-scale metabolic models provide a valuable framework to help interpret essentiality screens, since they both recapitulate knowledge on metabolic networks and allow prediction of gene essentiality under well-defined conditions. They have also allowed meaningful cross-validation of reconstructed metabolic networks with sets of gene essentiality results, providing insights on potential erroneous or incomplete metabolic knowledge, and on possible improvements [4,11,12]. In this article, we systematically exploit inconsistencies between model predictions and experimental results to improve a metabolic model reconstruction.

Our focus is on *Acinetobacter baylyi* ADP1, a strictly aerobic γ-proteobacterium. Although phylogenetically close to the *Acinetobacter baumanii* pathogenic strains, responsible for a growing number of nosocomial infections [13], *A. baylyi* ADP1 is an innocuous soil bacterium. Because of its metabolic versatility and high competency for natural genetic transformation, it is a model organism of choice for genetic and metabolic investigations [14-16]. As a soil bacterium, *A. baylyi* is able to degrade a wide range of molecules, including components of suberin, a protective polymer produced by plants in response to stress. Its harmlessness, nutritional versatility, and high capacity for adaptation have led bacteria of the *Acinetobacter* genus to be used for a variety of biotechnological applications– including the degradation of pollutants (e.g. biphenyl, phenol, benzoate, crude oil, nitriles) and the production of valuable biochemical products such as lipases, proteases, bioemulsifiers, cyanophycine and different kinds of biopolymers [17,18]. Following its sequencing and expert annotation [19], a genome-wide single-knockout mutant library was generated (ADP1 mutant collection [8]), enabling the high-throughput assessment of mutant phenotypes in defined growth conditions.

We report below on the reconstruction and refinement of a genome-scale metabolic model for *A. baylyi* with the help of high-throughput experimental data. Following an initial reconstruction using metabolic information extracted from the genome annotation and the literature, the model was iteratively assessed and improved by comparing its predictions with (1) large-scale growth phenotyping results of the wild-type strain on 190 distinct environments, (2) genome-wide gene essentiality data from the mutant collection, and (3) conditional gene essentiality data derived from growth phenotyping of *A. baylyi* mutants on eight defined media. We examined each inconsistency between experimental results and model predictions, and corrected the model when sufficient justifying evidence could be collected. Combining the three refinement steps, 1262 out of 1412 predictions were initially consistent with experimental results. Among the inconsistent cases, 65 led to improvements, increasing the completeness and accuracy of the model. The final version of the model, called iAbaylyi^{v4}, predicted accurately (1) 91% of the wild-type growth phenotypes, (2) 94% of the genome-wide gene essentialities, and (3) 94% of the phenotypic profiles of *A. baylyi* mutants on the tested media.

We developed a web interface which provides easy access to both model and experimental data. The interface allows browsing of the metabolic network, online computation of phenotype predictions, and comparison of predictions with experimental results [20].

Results and discussion

Initial model reconstruction

The genome scale model of *A. baylyi* was iteratively reconstructed following a process depicted in Figure 1. We first built an initial draft model iAbaylyi^{v1} using information from the genome annotation, metabolic pathways databases, and the literature. Although facilitated by the automated network reconstruction software PathoLogic [21],

Model versions	iAbaylyi v1 787 genes 859 reactions 697 metabolites		iAbaylyi v2 789 genes 873 reactions 702 metabolites		iAbaylyi v3 778 genes 874 reactions 701 metabolites		iAbaylyi v4 774 genes 875 reactions 701 metabolites
Experimental datasets	**Dataset 1** Growth phenotypes of wild-type strain on 190 carbon sources 1 strain x 190 environments			**Dataset 2** Genome-wide gene essentialities on succinate-supplemented minimal medium derived from ADP1 mutant collection buildup 3093 strains x 1 environment		**Dataset 3** Growth phenotypes of ADP1 mutant collection on 8 minimal media 2350 strains x 8 environments	
Model accuracy	**190**	**Total environments tested**	**190**	**767**	**Total genes tested**	**756**	**455**	**Total genes tested**	**452**
	164	total consistent	173	676	total consistent	712	422	total consistent	426
	26	total inconsistent	17	91	total inconsistent	44	33	total inconsistent	26
	86%	*global accuracy*	*91%*	*88%*	*global accuracy*	*94%*	*93%*	*global accuracy*	*94%*
	45	**Carbon sources**	**45**	**251**	**Essential genes**	**251**	**36**	**Conditionally essential genes**	**36**
	24	consistent	33	187	consistent	217	16	consistent	18
	21	inconsistent	12	64	inconsistent	34	20	inconsistent	18
	8	in model	1	*75%*	*accuracy*	*86%*	*44%*	*accuracy*	*50%*
	13	not in model	11						
	53%	*accuracy*	*73%*						
	145	**Non carbon sources**	**145**	**516**	**Dispensable genes**	**505**	**419**	**Dispensable genes**	**416**
	140	consistent	140	489	consistent	495	406	consistent	408
	31	in model	31	27	inconsistent	10	13	inconsistent	8
	109	not in model	109	*95%*	*accuracy*	*98%*	*97%*	*accuracy*	*98%*
	5	inconsistent	5						
	97%	*accuracy*	*97%*						
Model corrections	GPR	2 genes added		1 gene added ; 12 removed ; 36 GPR modified			4 genes removed ; 9 GPR modified		
	NETWORK	14 reactions added		4 reactions added ; 3 removed ; 4 modified			1 reaction added ; 1 modified		
	BIOMASS	-		2 biomass precursors added ; 4 removed			-		

Figure 1

***A. baylyi* metabolic model refinement process**. *A. baylyi* metabolic model was iteratively refined in three steps using datasets of experimental results. The initial reconstruction iAbaylyi^{v1} was assessed and improved using dataset 1; the resulting model iAbaylyi^{v2} was then assessed and refined using dataset 2, yielding iAbaylyi^{v3} which was again evaluated and refined using dataset 3, leading to the final model iAbaylyi^{v4}. Since only mutants corresponding to dispensable genes in dataset 2 could be phenotyped in dataset 3, gene essentialities revealed in dataset 3 are medium-specific, i.e. conditionally essential. Genes classified as conditionally essential in dataset 3 are conditionally essential on at least one environment. Genes classified as dispensable are dispensable on all tested environments. Model accuracy figures indicates for each dataset and its corresponding models the counts of consistent and inconsistent predictions. Accuracy is computed as the fraction of consistent predictions among all predictions. For dataset 1, Biolog results for metabolites that were not in the model were counted as consistent with predictions if the metabolite was not a carbon source and inconsistent if the metabolite was a carbon source. Model corrections figures summarize the corrections performed on each model component.

this initial reconstruction still required extensive manual curation (see Methods). The draft metabolic network generated by PathoLogic was first inspected to filter out and correct wrongly predicted pathways and reactions, and then completed by reviewing the expert genome annotations and the metabolic information contained in the literature. For instance, specific efforts were dedicated to properly include pathways accounting for the particular degradation capabilities of *A. baylyi*. Physiological information on *A. baylyi* was especially helpful to build the set of transport processes, as substrate specificities of transporters are difficult to deduce from genome annotation only. For each metabolite shown to be consumed by *A. baylyi* we added a corresponding transport reaction to the model. Out of 133 transporters, 23 were initially included in the model using this type of evidence only. The dependency between genes and reactions was modeled using Boolean rules, known as GPR (Gene-Protein-Reaction associations) [22]. These rules encode the presence of isozymes or enzymatic complexes for the catalysis of reactions, and predict the effect of genetic perturbations on the activity of reactions. GPR rules were first derived using homology with *E. coli* enzyme complexes [23] and then completed by manual curation. In order to model the metabolic and energetic demands associated with growth, we introduced a set of intermediary biomass reactions that synthesize generic cell constituents (e.g. protein, DNA, RNA, or lipid) from precursor metabolites, and a

global growth reaction consuming them in proportion defined by studies of biomass composition [24,25]. Energetic parameters required to predict quantitative growth rate using Flux Balance Analysis (FBA) were assumed to be similar to those of *E. coli* model (see Methods)[22]. No accurate measurement of *A. baylyi* growth yields could be used to validate these parameters, however. While such validation would be required to get more accurate predictions of growth yields, the current parameters already provide good approximate values (see Additional file 1 for a sensitivity analysis on these parameters). For the purpose of qualitatively predicting growth ability using Metabolite Producibility analysis (see Figure 2) [26], we designed a reduced list of biomass precursors which are all essential for growth in *in vitro* conditions. We used this list to predict qualitative growth phenotypes and compare them with those of phenotyping experiments on *in vitro* environments. *In vivo* environments may impose harsher conditions requiring additional metabolic responses; this list therefore represents a minimal set of essential precursors that may need to be expanded to properly predict growth phenotypes on more realistic environments [27]. The Methods section provides more details on the reconstruction process.

This initial reconstruction process led to the model iAbaylyi^{v1} gathering 859 reactions grouped in 7 metabolic categories and 697 distinct metabolites, 109 of which could be transported from the environment. As depicted in Figure 3, the model accounts for all main processes of *A. baylyi* metabolism, including biosynthetic routes, energy metabolism, and catabolic pathways. Genomic islands of catabolic diversity endow *A. baylyi* with the ability to degrade a wide variety of soil compounds [19]. The metabolic model reflects this nutritional versatility, as 20% of its reactions are dedicated to the catabolism of external compounds. A list of specific compounds that can be degraded by *A. baylyi* is provided in Table 1.

iAbaylyi^{v1} involves 787 genes out of the 1518 confirmed or putative enzymatic and transport genes of *A. baylyi*. A large majority (94%, 681/726) of the enzymatic reactions (excluding transporters) were associated with at least one gene, while the lower proportion (83%, 110/133) of transport reactions linked to genes is explained by the extensive use of physiological data to include them. The association of nearly all reactions with a gene confers a high reliability to the model. The few reactions that were introduced with no associated gene are most often supported by indirect evidence and introduced in order to fill gaps (See Additional file 2).

Most *A. baylyi* genes were annotated by expert curation; a third of the model genes relied on evidence conferring them a medium confidence level, e.g. limited homology with genes of known function, or conservation of amino acid motifs (Figure 4). While the evidence for these genes does not fully prove the existence of associated enzymatic activities, it suggests them with sufficient strength to justify adding the corresponding reactions in the model. The level of evidence of each gene was tracked for later use in interpreting inconsistent behaviors. Out of 262 reactions to which these genes contribute, 85 are solely catalyzed by medium-confidence genes, some of these being essential to the model viability. In addition, 35% of all coding sequences are still of unknown function in *A. baylyi*, and may leave gaps in the actual metabolic network. Integration of additional experimental data was thus crucial in order to validate the metabolic network and correct it when necessary.

Model validation and expansion using growth phenotype results

We used results of large-scale growth phenotyping experiments to perform a first round of model assessment and refinement. Using Biolog assays, we experimentally tested the wild-type strain ability to use 190 distinct metabolites as sole carbon and energy sources (see Methods). Using the model, we predicted the growth phenotypes of the wild-type strain on the corresponding *in silico* media and compared them to the experimental results.

Out of the 190 screened metabolites, 45 were found to be carbon and energy sources for *A. baylyi*. This relatively small fraction of carbon sources can be explained by the fact that Biolog microplates are only partially adapted to *A. baylyi*'s biotope: they feature sugars, nucleosides or amino acids but relatively few chemicals originating from plant compounds. iAbaylyi^{v1} model predicted 24 of them and missed 21 (see Figure 1). Eight of the missed carbon source metabolites were already present in the model, but with no associated transporter. Amongst them, seven would also be predicted as carbon and energy sources had the corresponding transporters been included. In order to resolve these inconsistencies, we added for each of them a generic transport reaction accounting for *A. baylyi*'s ability to utilize these compounds (see Table 2). Growth on the remaining metabolite (2-ketobutyrate) was contradicted by an additional individual growth experiment.

Thirteen carbon source metabolites were unknown to the metabolic model. For two of them, sorbate and tricarballylate, we were able to identify degradation pathways and add them to the model (see Table 2). Sorbate, an unsaturated fatty acid, can be degraded by fatty acids oxidation enzymes, which were already included in the model for the degradation of other fatty acids. Sorbate transport and degradation reactions were therefore added to the model using the same set of genes. Recently, genes coding for tricarballylate transport (*tcuC*), oxidation to cis-aconitate

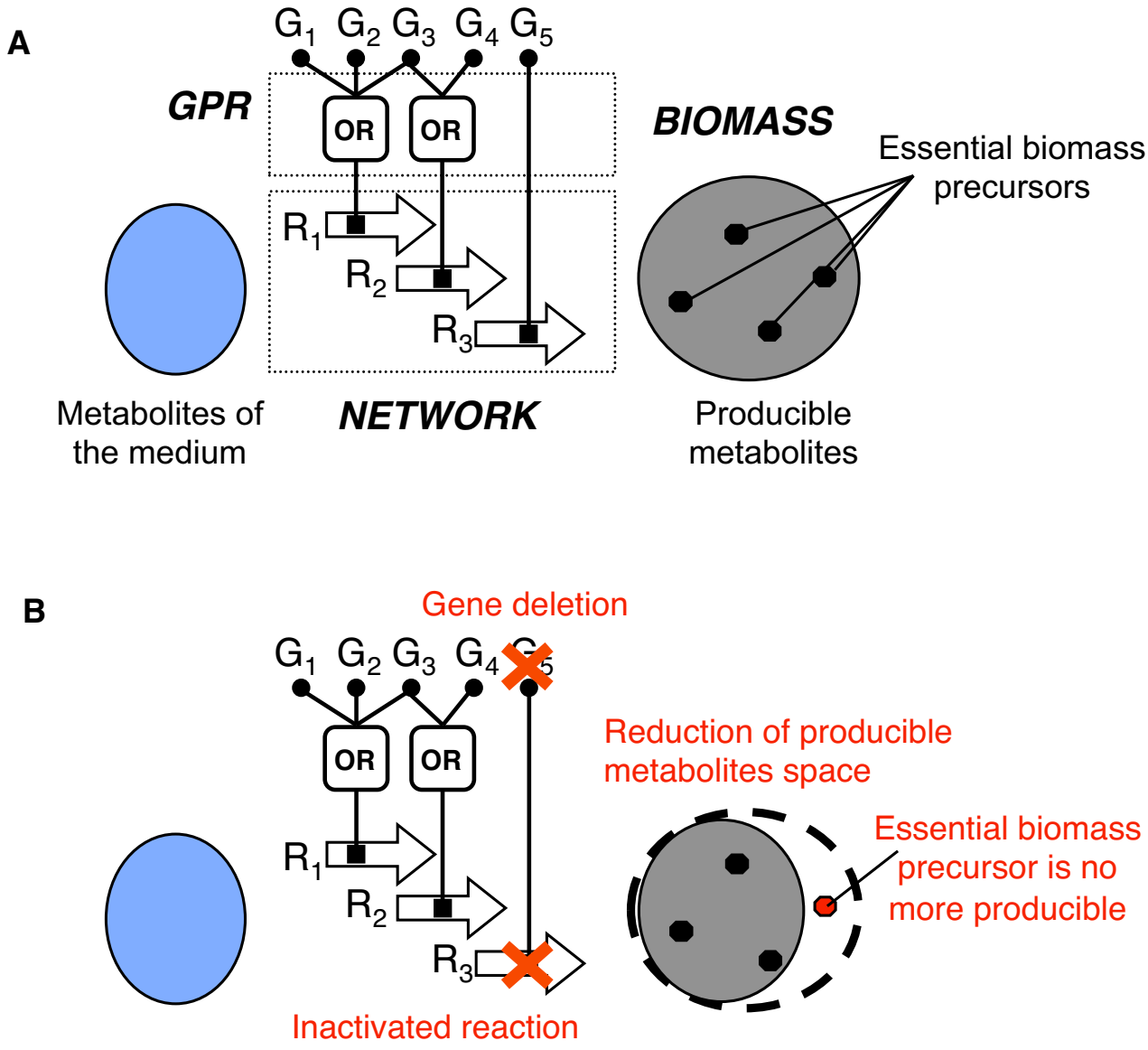


Figure 2
Modeling framework. (**A**) A metabolic model is represented as a combination of three model components: GPR Boolean rules associate genes (G_1 to G_5) with reactions (R_1 to R_3), the network of reactions defines the set of feasible biochemical transformations (illustrated by the arrows), and the set of essential biomass precursors defines the requirements for growth. Growth phenotypes are predicted by assessing whether all biomass precursors can be produced by the metabolic network from the set of metabolites from the medium [26] (see Methods) (**B**) Gene deletions potentially inactivate reactions, which in turn may reduce the space of producible metabolites. In case where a biomass precursor is no more producible, gene deletion is predicted lethal on the given medium.

(*tcuA* and *tcuB*), and for a regulatory protein required for *tcuABC* expression (*tcuR*) were identified in *Salmonella enterica* [28,29]. Highly homologous genes could be found in synteny in *A. baylyi*: ACIAD1536 (*tcuB*, 59% identity), ACIAD1537 (*tcuA*, 76% identity), ACIAD1541 (*tcuC*, 64% identity), ACIAD1539 (*tcuR*, 46% identity), and ACIAD1543 (*tcuR*, 44% identity). Following these clues, we expanded the model by implementing the corresponding transporter and degradation reaction, and annotated the corresponding genes. In four cases, dedicated growth experiments contradicted the Biolog result, weakening the case for further study (see Table 2). Finally,

Table 1: Some substrates involved in *A. baylyi* degradation pathways

Anthranillate	Octane
Benzoate	Straight chain dicarboxylic acids
Salicylate	Straight chain fatty acids
Catechol	Sarcosine
Chlorogenate	Propanaldoxime
Quinate	Propanenitrile
Shikimate	Propanamide
Coumarate	Malonate
Ferulate	Glucarate
Vanillate	Galactarate
Caffeate	Ethanesulfonate
Protocatechuate	

no relevant pathway could be found for the remaining seven unmodeled carbon sources. Further investigations are needed to identify the metabolic processes allowing *A. baylyi* to exploit these metabolites.

Conversely, only five of the 145 non-carbon source metabolites were wrongly predicted to be carbon sources by the model: 4-hydroxybenzoate, D-fructose, L-arginine, L-ornithine, and D-serine (see Figure 1 and Table 2). Experiments from [15] contradicted the Biolog result on 4-hydroxybenzoate, while additional individual experiments confirmed the Biolog results of the other four.

Interestingly, *A. baylyi* annotation describes a complete phosphotransferase (PTS) transport system for fructose (ACIAD1990 and ACIAD1993, *fruA* &*fruB*) coupled with a 1-phosphofructokinase (ACIAD1992, *fruK*) leading to fructose-1,6-bisphosphate (see Figure 5). In accordance with the annotation, the model predicts that fructose should be a carbon and energy source, yet this is not observed experimentally. To confirm the ability of the PTS system to transport fructose, we assessed experimentally the growth phenotype of the fructose bisphosphate aldolase (ACIAD1925, *fda*) knockout mutant (see Figure 5). The ΔACIAD1925 mutant could not be obtained on succinate-supplemented minimal media, reflecting the fact that Fda is required in the gluconeogenesis pathway to provide fructose-1,6-bisphosphate, an essential intermediate for building pentose-phosphates and polysaccharides. The mutant could however be obtained by adding fructose in the medium, showing that fructose could be

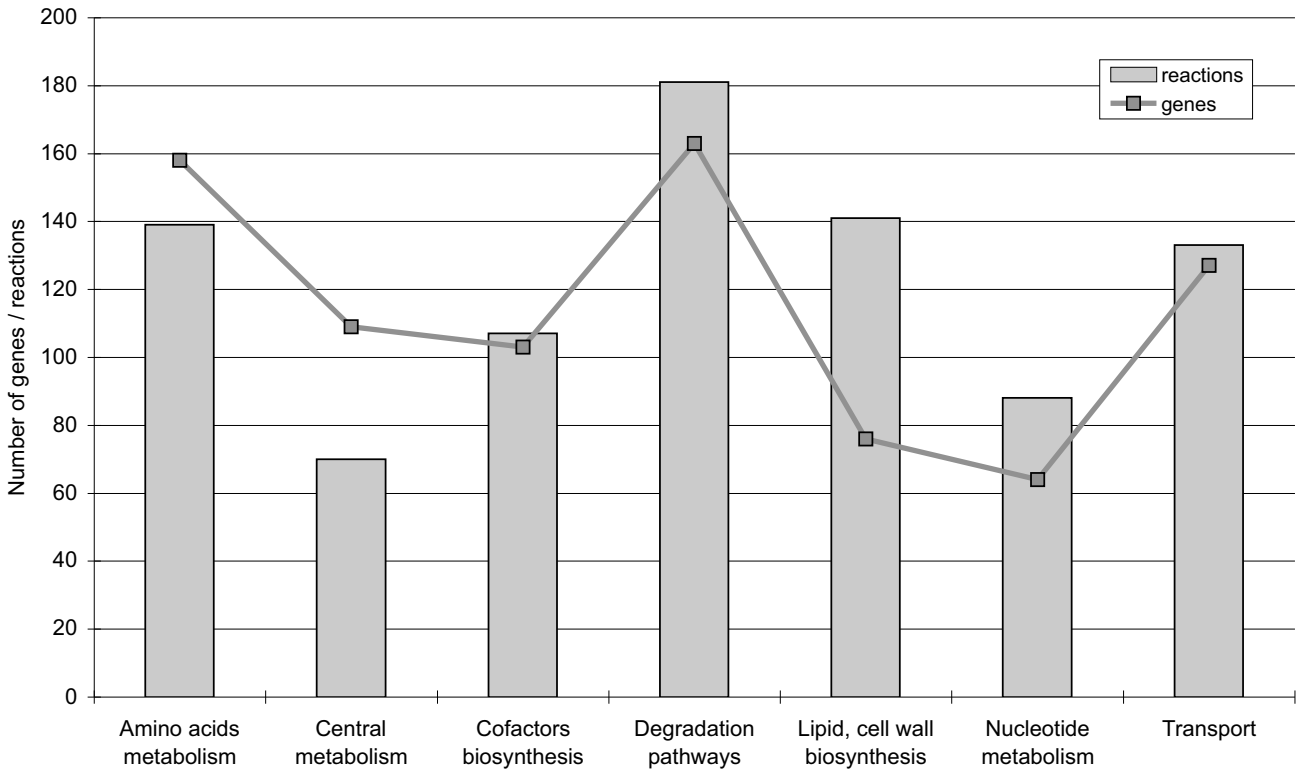


Figure 3
Number of reactions and genes in iAbaylyi^{v1} distributed by model metabolic categories. Reactions were associated with a unique metabolic category. Genes linked to several reactions may be associated with multiple categories.

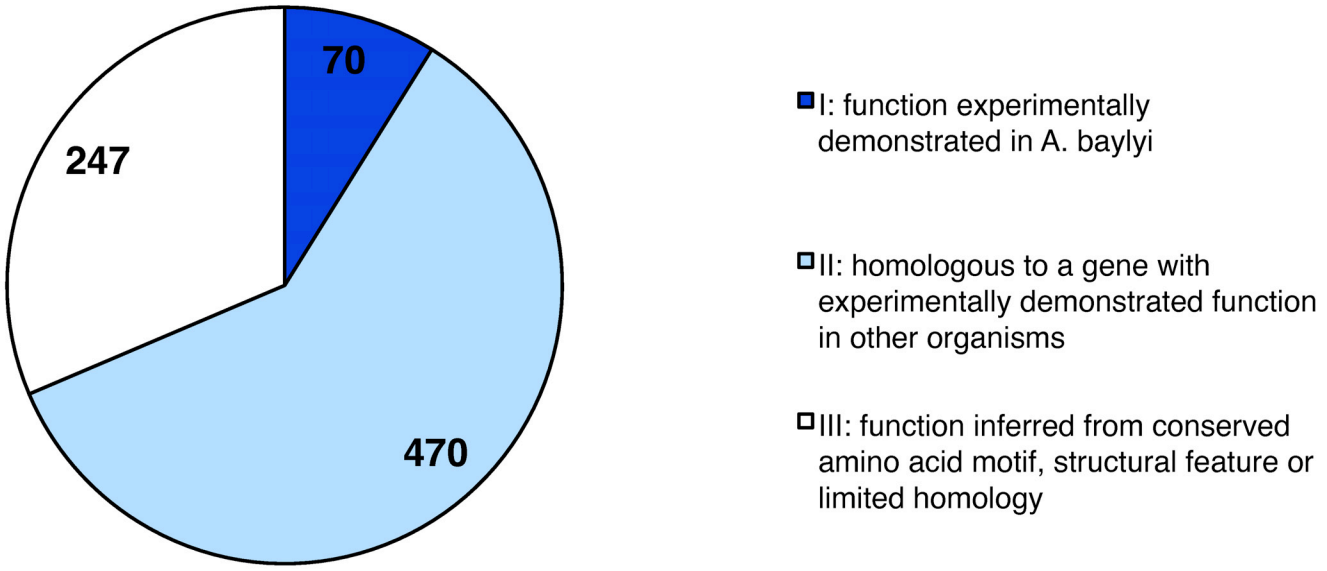


Figure 4
Distribution of annotation confidence levels for genes included in iAbaylyi^{v1} model. Confidence levels were assigned according to the type of evidence supporting gene annotation.

imported into the cell and converted to fructose-1,6-bisphosphate. The reason why *A. baylyi* is unable to use fructose as a sole carbon source remains yet to be investigated. Hypothetically, *A. baylyi* may be unable to use the Embden-Meyerhof-Parnas (EMP) pathway in the glycolytic direction, as it has been observed for the dissimilation of glucose [19,30].

As is the case in *E. coli*, L-ornithine and L-arginine are degraded by *A. baylyi* using the arginine succinyltransferase (AST) pathway. This pathway allows *E. coli* to use them as nitrogen sources, but not as carbon sources. Putative explanations include unsuitable regulation and inadequate transport [31]. Similar reasons may explain *A. baylyi*'s inability to use L-ornithine and L-arginine as carbon sources.

A. baylyi's genome annotation includes genes for D-serine transport (ACIAD0118 and ACIAD2662, *cycA*) and D-serine deaminase activity (ACIAD1048 *dsdA*), which should allow it to use D-serine as a carbon and nitrogen source. The interpretation of this inconsistency is also unclear; a similar unexplained inconsistency was pointed out in a study involving a metabolic model of *B. subtilis* [4].

Improvements to the model resulted in iAbaylyi^{v2}, raising predictive accuracy on Biolog-measured phenotypes from 86% to 91% of the growth phenotypes (see Figure 1). Detailed results of the comparison with Biolog results can be found in Additional file 3.

Systematic model improvement using gene essentiality data

In steps 2 and 3 of the model refinement process, we assessed and improved the model by comparing its predictions to experimentally determined gene essentialities, derived from the ADP1 mutant collection [8] (see Figure 1). Growth phenotypes of all single gene deletion mutants on the corresponding environments were predicted using metabolite producibility analysis (see Figure 2 and Methods). Predicted phenotypes were then compared to the genome-wide gene essentiality results in order to assess the accuracy of the model and to identify inconsistent predictions. Inconsistencies could be either *false essential* (genes falsely predicted essential by the model) or *false dispensable* (genes falsely predicted dispensable by the model) predictions. Since these inconsistencies are as many clues that the understanding of *A. baylyi*'s metabolism represented in the model is erroneous or incomplete, we examined them carefully in order to find interpretations and, when needed, refine the model.

We classified refinements into three categories according to the model component that was modified: GPR, NETWORK or BIOMASS (see Figure 2). These three components model different kinds of biological processes which contribute to determining the growth phenotype of mutant strains (see Methods). The GPR component, consisting of the GPR Boolean rules, computes the effect of the genetic perturbation on the activity of reactions in the model. The NETWORK component, the actual network of

Table 2: Biolog carbon sources inconsistently predicted by iAbaylyi^{v1} and corresponding corrections

Unpredicted Biolog carbon sources	**21**
Prediction corrected by addition of transporter	7
3-ketobutyrate	
butyrate	
D-aspartate	
L-asparagine	
L-glutamine	
propionate	
pyruvate	
Prediction corrected by addition of degradation pathway	2
sorbate	
tricarballylate	
Biolog result contradicted by additional experiment	5
2-ketobutyrate	
alpha-D-glucose	
D-malate	
D-xylose	
L-arabinose	
Uncorrected inconsistencies – no relevant pathway found	7
2-hydroxybutyrate	
bromo-succinate	
D-lactate methyl ester	
methylpyruvate	
tween 20	
tween 40	
tween 80	
Unpredicted Biolog non carbon sources	**5**
Biolog result contradicted by additional experiment	1
4-hydroxybenzoate *	
Uncorrected inconsistencies	4
D-fructose	
D-serine	
L-arginine	
L-ornithine	

* result from [15]. Numbers provide the count of inconsistencies pertaining to each category.

reactions, models the metabolic conversion capabilities of the organism. Finally, the BIOMASS component, consisting of the list of metabolites required for growth, models the biomass precursor requirements of the organism.

Model refinements

We performed two iterations of refinement using gene essentiality data (see Figure 1). In a first step, we used gene essentialities established during the construction of the ADP1 mutant library to derive an intermediary version of the model iAbaylyi^{v3}. This experimental dataset is nearly exhaustive as it covers 97% of all *A. baylyi* genes [8]. The mutant collection, built on succinate-supplemented minimal medium, revealed 499 essential genes for this medium. Half of these genes were present in the model (251/499), which is a significantly higher fraction than for all *A. baylyi* genes (24%, 789/3288). Although purely metabolic, the model thus already captured a large part of the bacterium's essential processes. The thoroughly curated but also purely metabolic *E. coli* model iAF1260 includes a similar proportion of *E. coli* essential genes on glucose-supplemented minimal medium (57%, 238/419) [12]. As shown in Figure 6, essential genes absent from the model were mainly related to functional categories lying outside of model scope, such as protein fate, DNA metabolism, transcription, or regulatory functions. On the other hand, essential genes involved in metabolic processes were largely covered by the model. iAbaylyi^{v2} already showed good agreement with the observed gene essentialities as 88% of the predictions were identical to the experimental results (respectively 95% of dispensable genes and 75% of essential genes present in the model, see Figure 1). As depicted in Figure 7, inconsistencies were homogeneously distributed across the metabolic categories of the model, with an exception for Transport and Degradation pathways, which gathered few inconsistencies. Genes in these categories are typically dedicated to the use of external substrates and most of them are not required for growth on succinate medium only. Their metabolic role could thus not be evaluated in this first experiment: most were accordingly both observed and predicted as dispensable. Gene essentiality experiments on a variety of media were needed to assess the functions of these genes in the appropriate environmental context.

It is worth noticing that inconsistency results support our choice to include medium-confidence genes into the model. Genes associated with medium-confidence metabolic annotations did not trigger more inconsistencies then high-confidence level genes. 18% (47/268) of reactions including at least one medium-confidence gene in their GPR are associated with an inconsistent gene, a similar proportion to that of reactions containing only high confidence genes (14%, 75/527). We examined the 91 inconsistent predictions of this step and refined the model for 47 of them (see Table 3 and below for details on the corrections). The refinements were implemented in iAbaylyi^{v3}, increasing global accuracy from 88% to 94%. Improvement was most noticeable for essential genes, as 86% were correctly predicted by iAbaylyi^{v3}. As discussed below, a high number of false isozymes, triggering *false dispensable* predictions, were detected in this refinement step.

In a second step, the model was evaluated against growth phenotyping assays of mutants from the ADP1 collection on 8 minimal media supplemented with varying carbon and nitrogen sources (see Table 4 and Methods). Since all *A. baylyi* mutants were first obtained on a succinate-supplemented minimal medium, essentialities revealed by these assays were strictly conditional. Furthermore, as the succinate-supplemented medium was already minimal,

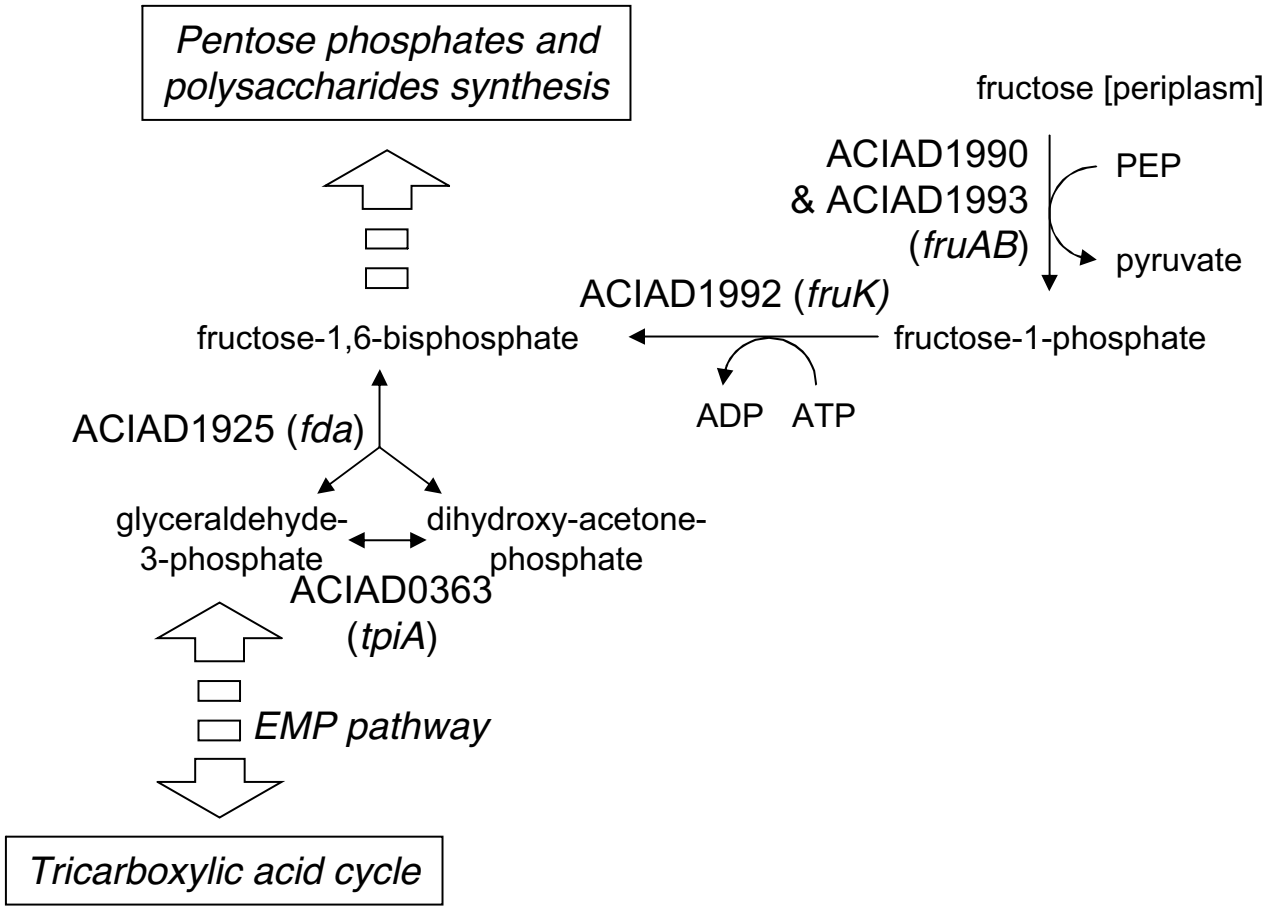


Figure 5
Map of fructose utilization pathway in *A. baylyi*. Fructose utilization pathway produces fructose-1,6-biphosphate which should be a precursor for the biosynthesis of pentose phosphates and polysaccharides and for the tricarboxylic acid cycle. Model accordingly predicts growth with fructose as sole carbon source. Phenotyping experiments show no growth of *A. baylyi* with fructose as sole carbon source. Supposedly, the Embden-Meyerhof-Parnas (EMP) pathway may not operate in the glycolytic direction in *A. baylyi*, as already observed for glucose utilization [19,30]. See main text for details.

the set of conditionally essential genes was restricted to the genes directly related to the use of the tested carbon and nitrogen sources. These were chosen to involve different parts of *A. baylyi* secondary metabolism (see Table 4). Overall, 455 knockout mutants corresponding to genes in the model could be phenotyped (see Figure 1).

Phenotyping experiments pointed out 2 to 10 conditionally essential genes (from the set of model genes) on each medium (Table 4). While a majority of these genes were essential on a single medium, some were found conditionally essential on several media. This revealed interdependencies between environments and might be related to processes specific to groups of environments. For instance, growth phenotypes on 2,3-butanediol and acetate exhibit similar characteristics since 2,3-butanediol is converted to acetate for its utilization [8]. The use of acetate as a carbon source requires the activation of the glyoxylate shunt, catalyzed by ACIAD1084 (isocitrate lyase) and ACIAD2335 (malate synthase G). These genes were therefore found to be essential on 2,3-butanediol and acetate only. Accordingly, the metabolic model correctly predicted the required use of this pathway and the subsequent essentiality of these genes on these media. As shown in Figure 1, iAbaylyi^{v3} accurately predicted the phenotypic profiles of 93% of all genes, leaving 33 genes with inconsistent predictions on at least one medium. Nine of them led to model corrections, again mainly in the GPR component of the model (see Table 3). These corrections, implemented in iAbaylyi^{v4}, slightly improved the predic-

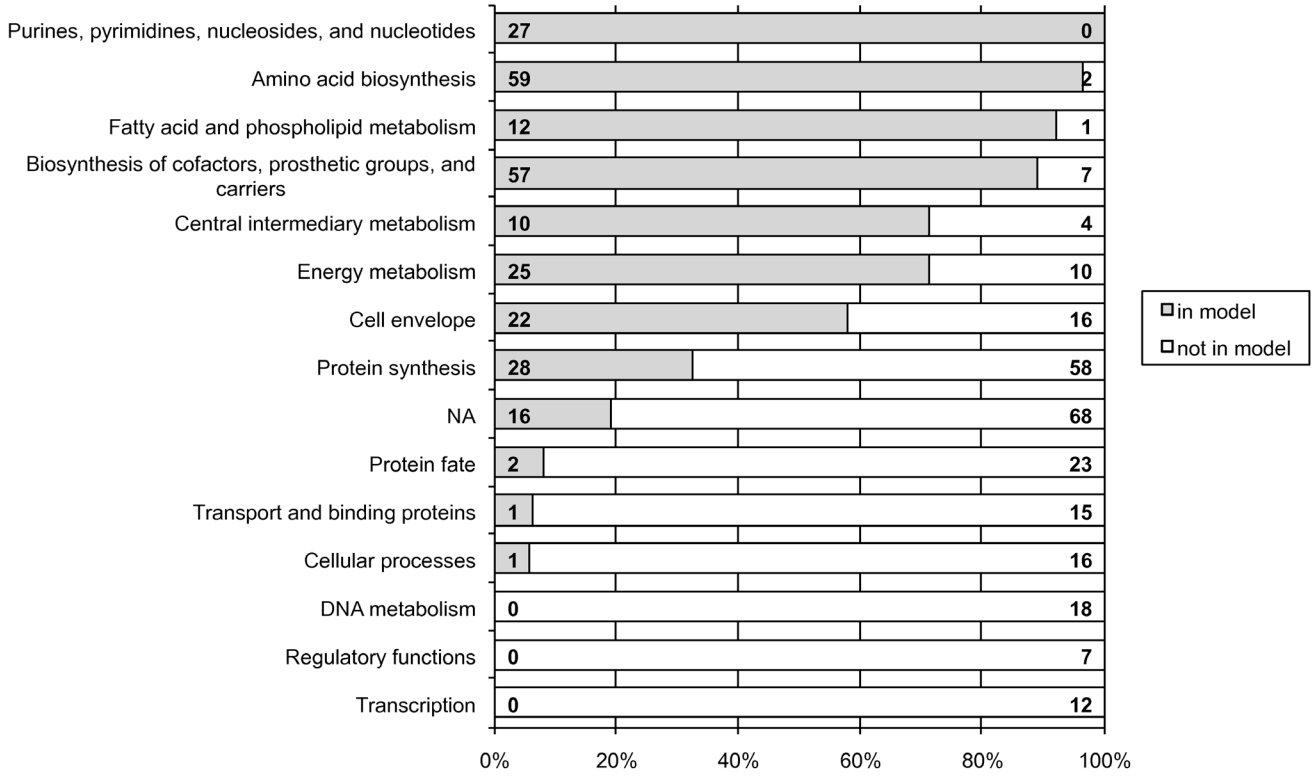


Figure 6

Proportion of *A. baylyi* essential genes covered by iAbaylyi^{v2} model distributed by TIGR role categories. TIGR role categories were obtained from TIGR automated annotation of *A. baylyi* [67]. Some genes were associated with multiple functional classes. NA: no TIGR role has been assigned. For each role category, absolute numbers of genes in the model (left) and not in the model (right) are provided.

tive accuracy for mutant phenotypes (94%) while keeping the predictive accuracy for the previous datasets unchanged.

Combining both refinement steps, 56 out of 124 inconsistencies led to model corrections. In the following sections, we will discuss these gene essentiality inconsistencies in more details irrespective to the dataset that triggered them (see also Table 3). Model corrections will be presented according to the model component that was modified.

GPR corrections

A majority of the model improvements (34/56) were applied to the GPR component, with a clear bias towards false dispensable inconsistencies: 26 GPR corrections pertained to experimentally essential genes against only 8 to experimentally dispensable genes (see Table 3). This large set of false dispensable predictions includes two main inconsistency types. In 22 cases, isofunctional genes with annotations of medium confidence were in fact unable to replace the activity of their deleted isozymes. For instance, ACIAD0964 and ACIAD2907 (*prs*) were identified in the initial reconstruction as isozymes for the catalysis of the ribose-phosphate diphosphokinase activity, which is required for the biosynthesis of 5-phosphoribosylpyrophosphate (PRPP) (see Figure 8A). The association of both genes to the activity relied on homologies with previously annotated genes in other organisms. The expected and predicted dispensability of ACIAD2907 was yet contradicted by its experimental essentiality. Looking further into the annotation evidence, ACIAD0964 function was supported by only limited homologies to previously known genes (second best hit after ACIAD2907 with *E. coli* gene *prsA*, with 25% identity). Conversely, ACIAD2907 function was supported by a stronger homology with *E. coli* gene *prsA* (68% identity) whose ribose-phosphate diphosphokinase has been experimentally confirmed [32]. The combination of the observed gene essentialities with the limited homology supporting the annotation of ACIAD0964 led us to correct the model by removing ACIAD0964 from ribose-phosphate diphosphokinase GPR. On the other hand, the functions of some isozymes with medium confidence level were corrobo-

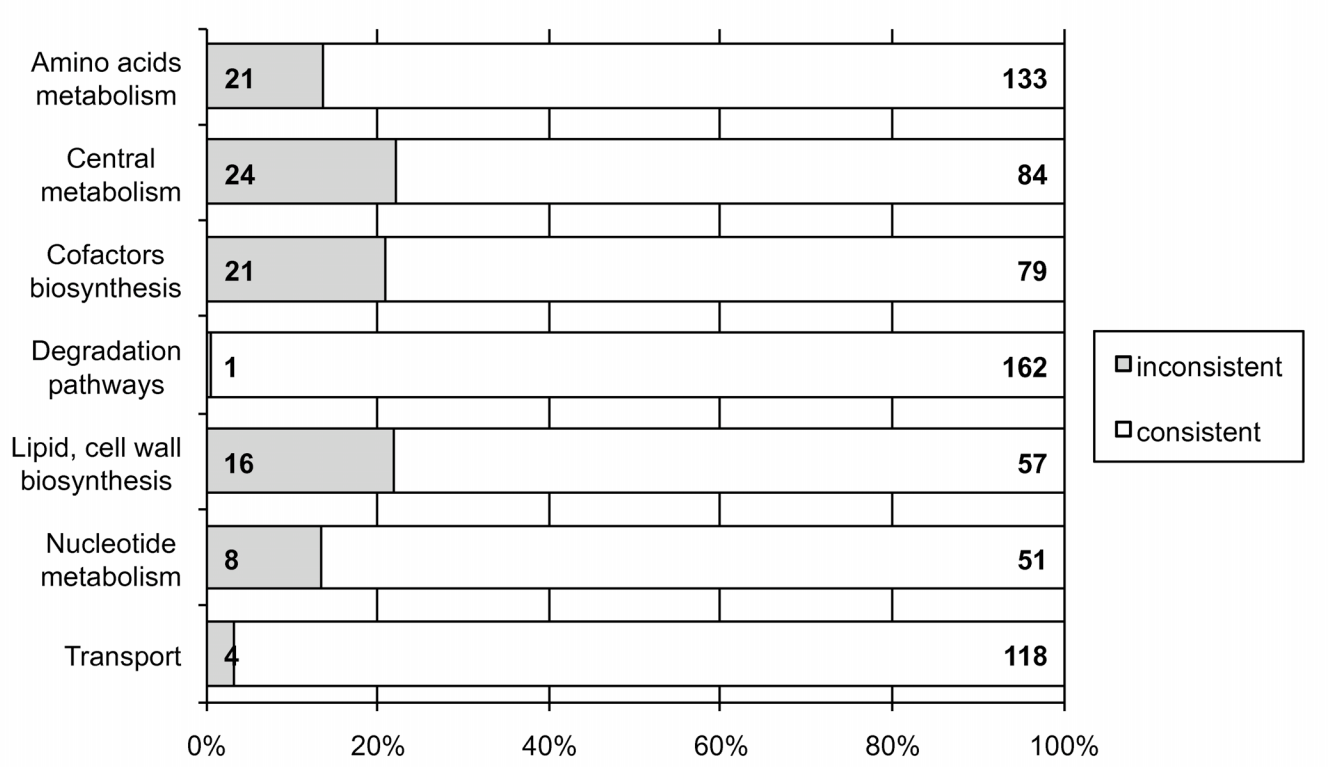


Figure 7
Consistency of gene essentiality predictions for dataset 2 and iAbaylyi^{v2} distributed by model metabolic categories. Proportions of genes having inconsistent predictions for essentiality on succinate-supplemented minimal medium in iAbaylyi^{v2} are shown for each model metabolic category. Genes linked to several reactions may be associated with multiple categories. For each metabolic category, absolute numbers of inconsistent (left) and consistent (right) gene essentiality predictions are provided.

rated by the gene essentialities. For instance, two isozymes were indirectly confirmed to have a dihydroxy-acid dehydratase activity, which is essential for the synthesis of valine, leucine and isoleucine. Two duplicate genes were associated with this activity: ACIAD1266 (*ilvD*) and ACIAD3636. While the annotation of ACIAD1266 is supported by a strong homology with *E. coli* gene *ilvD* (74% identity) whose activity has been experimentally shown [33], ACIAD3636's function was supported only by weaker homologies with the reference genes (37% identity with *E. coli* gene *ilvD*). Gene knock-outs revealed that both genes were dispensable while the essentiality of other genes in the pathway strongly suggested that the dihydroxy-acid dehydratase activity was required. This result strongly suggests that both genes could back up each other and therefore indirectly corroborates the functional assignment to ACIAD3636.

Further examination revealed that the duplicate genes are also found together in other organisms, including *Bradyrhizobium japonicum* and *Bordetella bronchiseptica*, and that *S. cerevisiae* possesses the gene ILV3, with a confirmed activity [34], which is homologous to ACIAD3636 (51% identity). Overall, amongst the reactions which were essential to iAbaylyi^{v2} viability and associated with an isozyme of medium confidence-level, 8 showed agreement between predictions and phenotypes while 11 triggered inconsistencies. In other words, while some medium-level genes were discarded thanks to essentiality data, a comparable fraction of genes was indirectly confirmed. This observation provides additional confirmation that essentiality data represents a valuable resource, as it helps validate or discard gene functions supported by reasonably good but non-conclusive evidence. It also provides an *a posteriori* validation of the usefulness of including medium-level annotations in the initial model, as failing to do so would have resulted in a significant loss of information in the *A. baylyi* metabolic model.

For three false dispensable predictions, we uncovered enzymatic complexes or functional dependencies between genes that were absent from the initial recon-

Table 3: Inconsistent gene essentiality predictions identified in refinement steps 2 and 3 and corresponding corrections and interpretations

CORRECTION			**56**
BIOMASS			**10**
biomass precursor not essential			9
ACIAD0076 (rmlB)	D	step 2	
ACIAD0078 (rmlD)	D	step 2	
ACIAD0079 (rmlA)	D	step 2	
ACIAD0080 (rmlC)	D	step 2	
ACIAD0086 (epsM)	D	step 2	
ACIAD0099 (galU)	D	step 2	
ACIAD0101 (pgi)	D	step 2	
ACIAD0104 (manB)	D	step 2	
ACIAD2429 (cyoE)	D	step 2	
missing essential biomass precursor			1
ACIAD1374 (ispU)	E	step 2	
GPR			**34**
activity simultaneously requiring all genes			3
ACIAD0661 (hisG)	E	step 2	
ACIAD1257 (hisZ)	E	step 2	
ACIAD3103 (ilvH)	E	step 2	
gene associated to another essential reaction			1
ACIAD2606	E	step 2	
isozyme not functional			22
ACIAD0151 (guaA)	E	step 2	
ACIAD0249 (ribC)	E	step 2	
ACIAD0871 (fabG)	E	step 2	
ACIAD1069 (lysS)	E	step 2	
ACIAD1255 (epd)	E	step 2	
ACIAD1323 (purF)	E	step 2	
ACIAD1375 (cdsA)	E	step 2	
ACIAD1736 (accC)	E	step 2	
ACIAD1737 (accB)	E	step 2	
ACIAD1925 (fda)	E	step 2	
ACIAD2227 (dctA)	E	step 2	
ACIAD2565 (gap)	E	step 2	
ACIAD2666	E	step 2	
ACIAD2907 (prs)	E	step 2	
ACIAD3062 (folK)	E	step 2	
ACIAD3249 (ribA)	E	step 2	
ACIAD3365 (murE)	E	step 2	
ACIAD3371 (gltX)	E	step 2	
ACIAD1710 (pcaC)	E	step 3	
ACIAD2018 (ald1)	E	step 3	

NO CORRECTION			**68**
Validated explanation			**6**
experimental error			1
ACIAD0108 (lldD)	D	step 3	
known gap in the understanding of pathway			4
ACIAD0856 (bioA)	E	step 2	
ACIAD0857 (bioF)	E	step 2	
ACIAD0859 (bioD)	E	step 2	
ACIAD2045 (bioB)	E	step 2	
unmodeled auxotrophy			1
ACIAD3523 (metE)	E	step 2	
Hypothetical explanation			**32**
ACIAD0178 (atpI)	E	step 2	
ACIAD0180 (atpB)	E	step 2	
ACIAD0182 (atpE)	E	step 2	
ACIAD0183 (atpF)	E	step 2	
ACIAD0184 (atpH)	E	step 2	
ACIAD0185 (atpA)	E	step 2	
ACIAD0186 (atpG)	E	step 2	
ACIAD0187 (atpD)	E	step 2	
ACIAD0188 (atpC)	E	step 2	
ACIAD0556 (ndk)	D	step 2	
ACIAD0650 (argJ)	E	step 2	
ACIAD1150 (pyrC)	E	step 2	
ACIAD1346 (sodB)	E	step 2	
ACIAD1358 (rpiA)	E	step 2	
ACIAD2282 (sahH)	D	step 2	
ACIAD2314 (metZ)	E	step 2	
ACIAD2458 (glnA)	E	step 2	
ACIAD2842 (pckG)	E	step 2	
ACIAD2847 (folD)	E	step 2	
ACIAD3155 (mdh)	E	step 2	
ACIAD3349 (gltD)	E	step 2	
ACIAD3350 (gltB)	E	step 2	
ACIAD3470 (msuE)	E	step 2	
ACIAD3506 (aceF)	E	step 2	
ACIAD0101 (pgi)	E	step 3	
ACIAD0546	E	step 3	
ACIAD0556 (ndk)	D	step 3	
ACIAD1021	D	step 3	
ACIAD1707 (pcaB)	E	step 3	
ACIAD1711 (pcaH)	E	step 3	

Table 3: Inconsistent gene essentiality predictions identified in refinement steps 2 and 3 and corresponding corrections and interpretations *(Continued)*

ACIAD2088 (aspQ)	E	step 3	
ACIAD2983 (gcd)	E	step 3	
presence of an alternate enzyme			6
ACIAD1231 (argD)	D	step 2	
ACIAD1642 (uppP)	D	step 2	
ACIAD2968 (ispA)	D	step 2	
ACIAD1020 (acoL)	D	step 3	
ACIAD1715 (quiX)	D	step 3	
ACIAD2984	D	step 3	
spontaneously occurring reaction			1
ACIAD2819	D	step 3	
wrong complex subunit			1
ACIAD0799	D	step 2	
NETWORK			**12**
false alternate pathway in the model			7
ACIAD0239 (ppa)	E	step 2	
ACIAD0547 (proA)	E	step 2	
ACIAD1105 (adk)	E	step 2	
ACIAD1920 (glnS)	E	step 2	
ACIAD2560 (proB)	E	step 2	
ACIAD3032 (proC)	E	step 2	
ACIAD0901 (dut)	E	step 2	
missing alternate pathway in the model			5
ACIAD0106 (lldP)	D	step 2	
ACIAD0451 (katA)	D	step 2	
ACIAD0930 (glpK)	D	step 2	
ACIAD1045 (metH)	D	step 2	
ACIAD0106 (lldP)	D	step 3	
ACIAD1712 (pcaG)	E	step 3	
ACIAD1744 (aspA)	E	step 3	
No precise interpretation			**30**
ACIAD0072 (ugd)	E	step 2	
ACIAD0173 (rhtB)	E	step 2	
ACIAD0382 (ubiB)	D	step 2	
ACIAD0505 (purU1)	E	step 2	
ACIAD1482 (kdsD)	D	step 2	
ACIAD1483 (kdsC)	D	step 2	
ACIAD2283 (metF)	D	step 2	
ACIAD2290 (cydA)	E	step 2	
ACIAD2525	E	step 2	
ACIAD2667 (pdxB)	D	step 2	
ACIAD2788	E	step 2	
ACIAD2880 (sdhA)	D	step 2	
ACIAD2911 (panD)	D	step 2	
ACIAD3503 (guaB)	E	step 2	
ACIAD3510 (lpxC)	D	step 2	
ACIAD0086 (epsM)	E	step 3	
ACIAD0335 (fadB)	E	step 3	
ACIAD0382 (ubiB)	D	step 3	
ACIAD0922	E	step 3	
ACIAD2070 (metI)	E	step 3	
ACIAD2282 (sahH)	D	step 3	
ACIAD2283 (metF)	D	step 3	
ACIAD2667 (pdxB)	D	step 3	
ACIAD2755	E	step 3	
ACIAD2875 (sucB)	E	step 3	
ACIAD2876 (sucA)	E	step 3	
ACIAD2880 (sdhA)	D	step 3	
ACIAD2911 (panD)	E	step 3	
ACIAD3071 (cysM)	E	step 3	
ACIAD3549 (gshA)	E	step 3	

Inconsistencies identified during the refinement steps using mutant library essentialities (step 2) and mutant growth phenotypes on 8 media (step 3). Inconsistencies leading to corrections (left column) are listed according to the model component that was corrected: GPR, NETWORK, and BIOMASS. Inconsistencies with no correction (right column) are listed according to the level of interpretation that could be drawn. Numbers provide the count of inconsistencies pertaining to each correction or interpretation category. For each inconsistency, E or D indicates the experimental phenotype of the mutant: E: gene is essential (on at least one medium for step 3), D gene is dispensable (on all media for step 3).

Table 4: Mutant phenotyping experiments: growth media and experimental results for genes included in iAbaylyi^{v3}

Source of[1]		Essentiality		Specific metabolic pathways *a priori* involved
Carbon	**nitrogen**	**E**	**D**	
acetate	ammonia	5	431	Glyoxylate shunt
L-asparagine	ammonia	3	445	Asparagine and aspartate degradation
D-2,3-butanediol	ammonia	10	433	Butanediol to acetoin to acetyl-coa degradation, glyoxylate shunt
D-glucarate	ammonia	5	413	Glucarate to 2-oxoglutarate degradation
β-D-glucose	ammonia	7	432	Entner-Doudoroff pathway
L-lactate	ammonia	2	445	Lactate dehydrogenase
quinate	ammonia	8	436	Quinate to protocatechuate to acetyl-coa and succinyl-coa degradation
succinate	*urea*	3	442	Urease

[1] *Italic text* indicates the changed carbon or nitrogen source with respect to the medium used for mutant construction (succinate and ammonia).
E: number of conditionally essential genes
D: number of dispensable genes

struction: genes thought to be isozymes were in fact jointly required to catalyze the reactions. As an illustration, ACIAD0661(*hisG*) and ACIAD1257 (*hisZ*) were initially assigned as isozymes of ATP phosphoribosyltransferase reaction in the pathway of histidine biosynthesis (see Figure 8A). The observed essentiality of both genes suggested that they were both necessary to the activity. Further examination of the literature confirmed that, unlike in *E. coli*, ACIAD0661 forms a complex with ACIAD1257 [35]. In *E. coli*, *hisG* codes for an enzyme that is active on its own and is not part of a complex.

Amongst the false essential predictions which led to modifications of the GPR component, six cases involved associating additional enzymes to reactions. For instance, ACIAD2968 (*ispA*, farnesyl diphosphate synthase) was observed to be dispensable, even though it is the only catalyst of two reactions essential for the biosynthesis of isoprenoids, which are the precursors of vital cofactors (see Figure 8B). Previous work showed for *E. coli* that *ispA* was dispensable and that *ispB* (octaprenyl diphosphate synthase) and *ispU* (undecaprenyl diphosphate synthase) could perform these activities [36]. *A. baylyi*'s homologues to these genes – ACIAD2940 (*ispB*) and ACIAD1374 (*ispU*) – were therefore added as isozymes of ACIAD2968 for both reactions (see Figure 8B).

The remaining types of GPR refinement involved associating genes with already existing essential reactions (ACIAD2606: associated with nicotinate-nucleotide adenylyltransferase activity, which is essential for NAD biosynthesis), adding new complex subunits (ACIAD0799: falsely considered as a sulfite reductase subunit and replaced by ACIAD2981 after further investigations) or assigning spontaneous activity (ACIAD2819: encodes for gluconolactonase activity which has been shown to occur spontaneously [37]). See Additional file 3 for further details on these corrections.

NETWORK corrections

Twelve gene essentiality inconsistencies from datasets 2 and 3 led us to improve the NETWORK component of the model (see Table 3). Two types of inconsistencies fall within this category. On the one hand, false dispensable predictions may indicate that alternate pathways present in the model are either inactive for the experimental conditions under observation or not present at all. Seven discrepant predictions led us to reconsider alternate pathways in the model. For instance, ACIAD0822, ACIAD0823, and ACIAD0824 (*gatABC*), annotated as aspartyl/glutamyl-tRNA amidotransferase, catalyzed in iAbaylyi^{v2} the synthesis of charged glutamine-tRNA and charged asparagine-tRNA through the transamidation of misacylated glutamate-tRNA(Gln) and aspartate-tRNA(Asn) (see Figure 8C). Charged glutamine-tRNA can also be produced by the direct charging of glutamine on its tRNA using the glutaminyl-tRNA synthetase enzyme (ACIAD1920, *glnS*), however. The observed essentiality of ACIAD1920 is inconsistent with the redundancy of these two pathways, suggesting that the transamidation of glutamate-tRNA(Gln) does not occur *in vivo*. Furthermore, aspartate-tRNA(asn) transamidation is actually the only way of producing asparagine, as *A. baylyi* is lacking both asparagine synthetase and asparaginyl-tRNA synthetase enzymes. This result strongly suggests that, in *A. baylyi*, ACIAD0822-0824 genes are predominantly employed for asparagine-tRNA synthesis. To account for ACIAD1920 essentiality, we thus removed the glutamate-tRNA(Gln) transamidation pathway from the metabolic network.

On the other hand, false essential predictions may suggest that alternate pathways are missing from the model. Corrections of this type involve searching for new metabolic

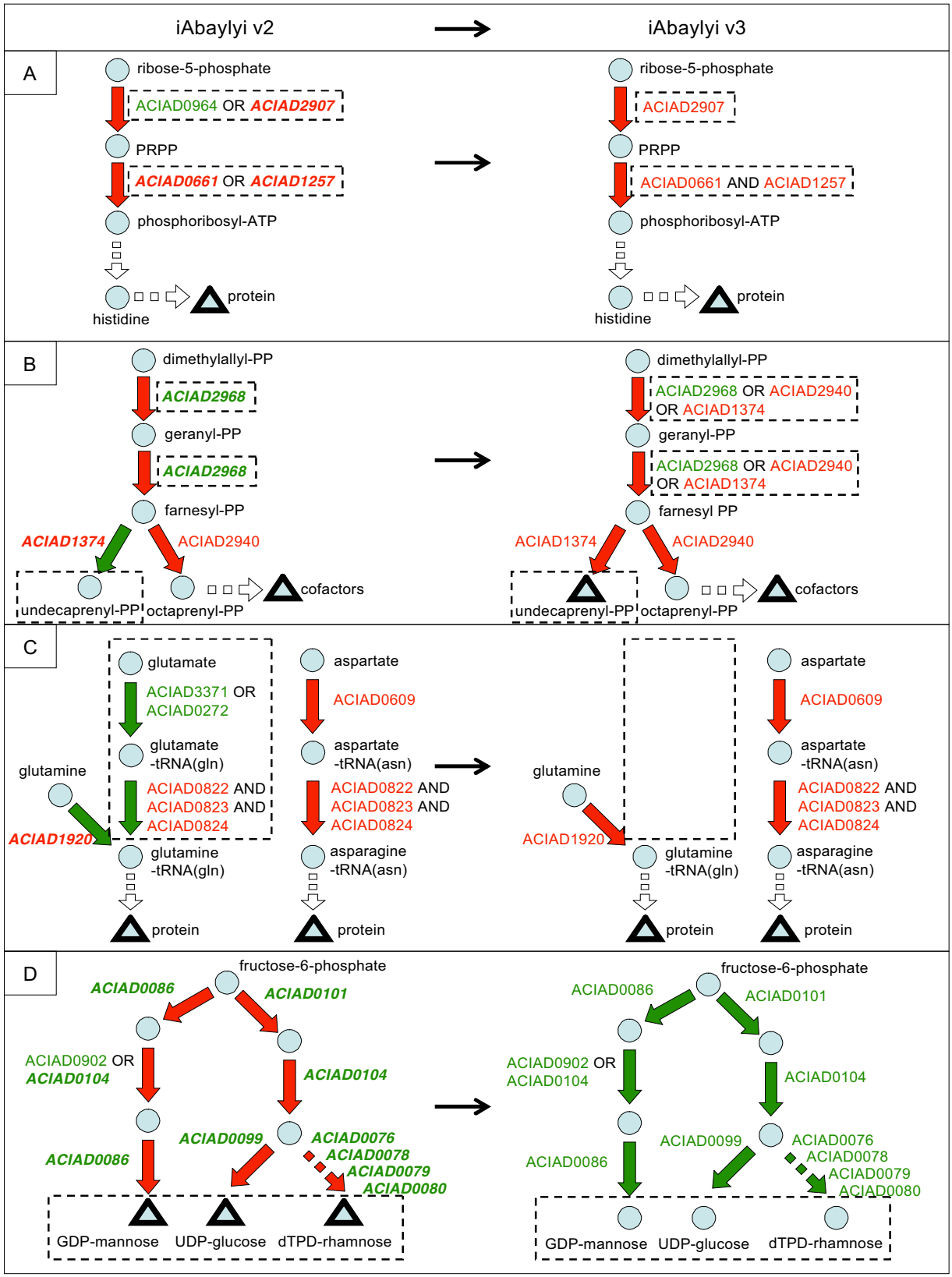


Figure 8 *(see legend on next page)*

Figure 8 *(see previous page)*
Model correction examples. Examples of model corrections performed between iAbaylyi^{v2} (left) and iAbaylyi^{v3} (right) models. Metabolites are depicted by blue circles and triangles, triangles indicating essential biomass precursors. Reactions are represented by arrows colored in red if they are predicted essential and in green if they are predicted dispensable. Gene names are indicated next to reaction arrows; they are written in red if they are experimentally essential and in green if they are dispensable. Genes with inconsistent predictions are written in bold italic. Dashed boxes indicate components that have been modified. Further evidence for model corrections are shown in main text and Additional file 3. (**A**) First steps of histidine biosynthesis. Unpredicted essentiality of ACIAD2907 encoding for ribose-phosphate diphosphokinase activity was corrected by removing the alternate gene ACIAD0964 from the reaction GPR. Unpredicted essentialities of ACIAD0661 and ACIAD1257, catalyzing the ATP phosphoribosyltransferase reaction, were corrected by assigning them as complex subunits instead of isozymes in the reaction GPR. (**B**) Isoprenoids biosynthesis. Unpredicted dispensability of ACIAD2968, catalyzing farnesyl-diphosphate and geranyl-diphosphate synthases activities, was corrected by adding ACIAD1374 (undecaprenyl-diphosphate synthase) and ACIAD2940 (octaprenyl-diphosphate synthase) as isozymes. Unpredicted essentiality of ACIAD1374 was resolved by adding undecaprenyl-PP to the set of essential biomass precursors. (**C**) Synthesis of charged glutamine-tRNA(gln) and asparagine-tRNA(asn). Unpredicted essentiality of ACIAD1920, encoding for glutaminyl-tRNA synthetase activity, was corrected by removing from the model the alternate pathway using aspartyl/glutamyl-tRNA amidotransferase enzyme (ACIAD0822-0824). (**D**) Biosynthesis of polysaccharides. Unpredicted dispensabilities of all genes involved in GDP-mannose, UDP-glucose, and dTDP-rhamnose synthesis were corrected by removing these three metabolites from the list of essential biomass precursors.

activities, a task that is open-ended and exploratory in nature and is likely to require additional experimental work. Five inconsistencies led to the addition of new reactions to the model, mainly for the transport of metabolites.

BIOMASS corrections

Ten inconsistent gene essentiality predictions led to modifications of the BIOMASS component (see Table 3). False essential inconsistencies can reveal biomass precursors that are not necessary to the viability of the cell on the tested environments, yet commonly produced by the wild-type strain. For instance, a large fraction of the BIOMASS modifications (8/10) were found in the biosynthesis of polysaccharides. Based on studies of the lipopolysaccharides composition of *Acinetobacter* species [38,39], three nucleotide sugars were initially included in the list of essential biomass precursors. All genes specifically involved in the synthesis of these sugars were found to be dispensable for growth on these *in vitro* environments (see Figure 8D). Further investigations are needed to analyze the composition of polysaccharides in the corresponding mutants and interpret the robustness to these deletions. Although dispensable in our experimental growth conditions, complete polysaccharides are likely to be essential on more realistic environments. Cell surface polysaccharides play an important role to help colonization and prevent desiccation while secreted polysaccharides are assumed to provide *A. baylyi* with better uptake capabilities of hydrophobic compounds in natural environments [19,40]. In order to account for these viable phenotypes on our experimental conditions, all three sugars were removed from the list of biomass precursors. Conversely, false dispensable inconsistencies may uncover essential metabolites that were initially overlooked. For instance, undecaprenyl diphosphate, a cofactor required for the synthesis of peptidoglycan, was not part of the biomass precursors list in iAbaylyi^{v2}. ACIAD1374 (*ispU*, undecaprenyl pyrophosphate synthetase), involved in its synthesis, was observed essential, although predicted dispensable (see Figure 8B). As this cofactor is regenerated during the peptidoglycan building process, its synthesis was actually not required at steady state. We therefore added undecaprenyl diphosphate to the list of essential metabolites in order to account for its required synthesis and resolve the unpredicted essentiality of ACIAD1374. An alternate method was recently introduced to account for the non-constitutive requirement for cofactors [27]. Small consumption terms are added for each cofactor in the equation of reactions involving them, thereby creating a replenishing flux of cofactor when reactions are active. This replenishing flux enforces the synthesis of the cofactor when required. While this method allows discarding cofactors from the general biomass requirements, it involves remodeling the reaction equations in an artificial manner.

Interpretation of remaining inconsistencies

The analysis of inconsistent predictions did not always lead to model refinement. Either the explanation of the discrepancy did not lead to model refinement, or no explanation interpreting the discrepancy could be validated.

Six discrepancies were confidently interpreted yet did not lead to model modifications (see Table 3). In one case, we identified a wrong experimental result. Four inconsistencies pertained to the pathway of biotin synthesis, whose

essentiality could not be accounted for by the model. Since the initial step of this pathway is unknown, it could not be linked to the metabolic network, preventing the model from simulating biotin synthesis. One inconsistency was caused by a requirement for a cofactor that could not be modeled. Two different methionine synthase enzymes catalyze the conversion of homocysteine to methionine: one B12-independent encoded by ACIAD3523 (*metE*) and one B12-dependent encoded by ACIAD1045 (*metH*). Since coenzyme-B12 is neither synthesized by *A. baylyi* nor provided in the experimental media, the ΔACIAD3523 mutant was unable to use the MetH enzyme to synthesize methionine. The model could not account for this B12 auxotrophy of the ΔACIAD3523 mutant. In order to properly account for the dependency between MetH activity and the presence of a cofactor, the replenishing flux method can be employed [27] or the modeling framework could be extended by introducing rules that state which conditions are required for the enzymes to be active. The introduction of this additional layer of rules has already been proposed to account for regulatory constraints [41] and may be helpful to explain a number of inconsistent phenotypes.

For 62 inconsistencies, we could not reach a validated explanation within the scope of this global analysis (see Table 3). For 32 of them, we could formulate hypothetical interpretations, all of which need experimental confirmation. A high proportion of these possible interpretations involve regulatory processes. For instance, *A. baylyi* possesses like *E. coli* two distinct enzymes for glutamate synthesis: glutamate synthase, encoded by ACIAD3350 (*gltB*) and ACIAD3349(*gltD*), and glutamate dehydrogenase, encoded by ACIAD1110 (*gdhA*). In *E. coli*, these pathways were shown to be regulated in response to nitrogen limitations [42]: glutamate synthase is used at low ammonium concentrations while glutamate dehydrogenase is used at high ammonium concentrations. *E. coli* strains lacking glutamate synthase show severe growth deficiency at low ammonium concentrations [42]. Similarly, ACIAD3350 and ACIAD3349 were found essential in *A. baylyi* on the succinate-supplemented minimal medium. These phenotypes contradicted model predictions, which considered the alternate pathway for glutamate synthesis. Further investigation would be required to fully understand the regulatory processes at work in this pathway for *A. baylyi* and extension of the modeling framework should be conducted to account for regulatory processes within the model.

The remaining 30 inconsistencies could not be given a clear interpretation and also require further investigations.

The final model: iAbaylyi^{v4}

The overall refinement process led to the final model iAbaylyi^{v4} gathering 774 genes, 875 reactions and 701 metabolites (see Figure 1). iAbaylyi^{v4} integrates all refinements resulting from the three experimental datasets introduced in this work. Accordingly, its predictions are consistent with the experimental results in 91% of the cases for dataset 1, 94% of the cases for dataset 2, and 94% of the cases for dataset 3. Compared with iAbaylyi^{v1}, it was expanded by 19 reactions and 2 genes, while 3 reactions and 16 genes were removed in the refinement process (see Figure 1, Model corrections).

An online software tool for the exploration of Acinetobacter baylyi *metabolism*

In order to facilitate the exploration of *A. baylyi* metabolism using the genome scale model, we created NemoStudio [20] (Combe *et al*, in preparation), a web interface combining a simulation layer for the model with AcinetoCyc, *A. baylyi* Pathway-Genome Database [21]. NemoStudio gathers data on functional genomics annotations, metabolic reactions and pathways, and experimental mutant phenotyping results within a single interface. Additionally, it allows performing phenotype predictions using the constraint-based model.

AcinetoCyc gathers information on the metabolic network of *A. baylyi* and is used to display interactive metabolic maps. After its initial automated construction using PathoLogic [21], AcinetoCyc has been undergoing constant curation. It includes all metabolic reactions present in the model.

NemoStudio integrates the latest version of *A. baylyi* metabolic model, iAbaylyi^{v4}. Growth phenotype predictions can be performed for any set of environmental conditions and genetic perturbations of this study. We implemented both Flux Balance Analysis (FBA) and Metabolite Producibility methods to predict growth phenotypes (see Methods). When performed on sets of environmental conditions and sets of gene deletions, prediction results are displayed in a table format in parallel to the actual experimental results. Predictions can thus be readily compared with the experimental observations. Furthermore, predicted and experimental phenotypes are both displayed on AcinetoCyc metabolic maps, and conversely gene deletions can be directly set from these metabolic maps (see Figure 9). When performed for a single environment and a single genetic perturbation, FBA predicts an optimal flux distribution towards biomass production; these fluxes are both displayed in a table and on AcinetoCyc metabolic pathways.

The availability of this resource as a web interface makes it easily usable by scientists interested in *A. baylyi* metabo-

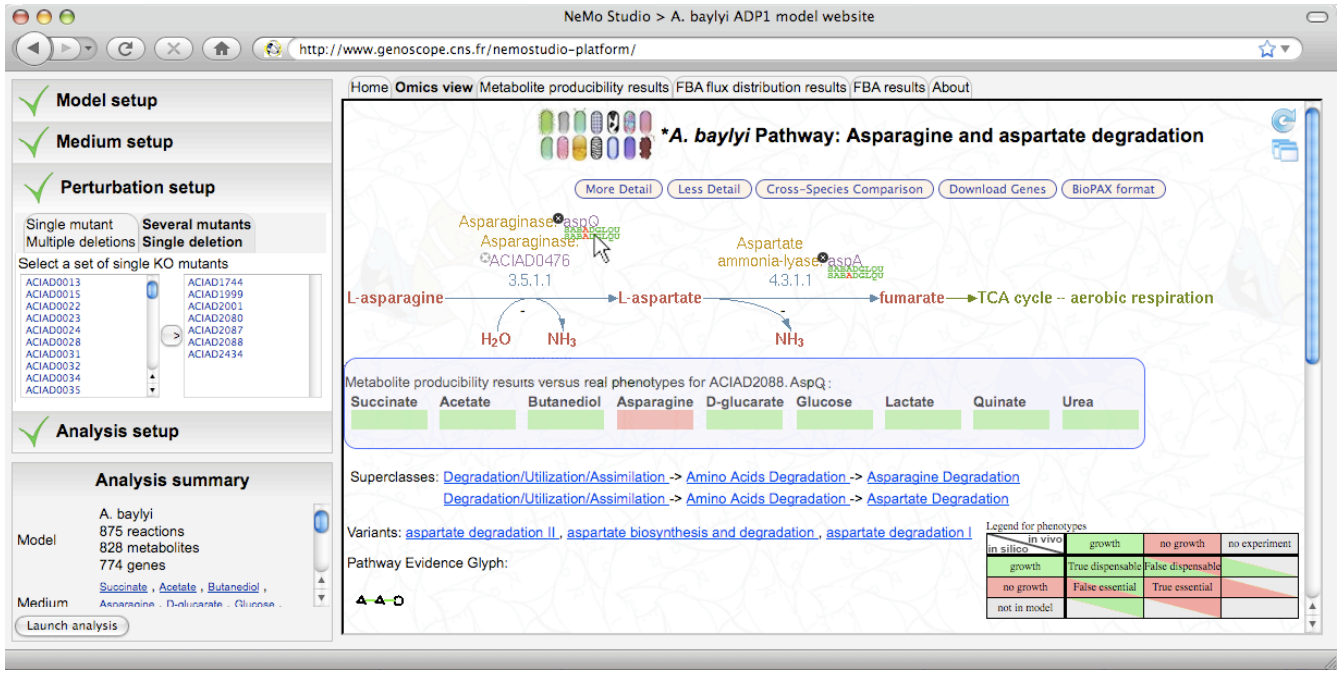


Figure 9

Screenshot of NemoStudio web interface. The web interface is divided in two parts. The left panel is dedicated to setting the analyses performed on the metabolic model. Simulated media, gene knockouts and type of analysis (metabolite producibility or flux balance analysis, see Methods) can be set in this panel. The right panel displays results in various formats for the selected type of analysis. The "omics view" part maps the predicted and experimental results on AcinetoCyc metabolic maps.

lism. Compared with previous web-based software for genome-scale metabolic modeling [27], the *A. baylyi* NemoStudio interface provides better interactivity, direct visualization of results on metabolic maps and integrated comparison with experimental data. By interfacing as much as possible results deriving from systems level analyses with experimental data of various forms, it allows the simultaneous exploitation of both information types.

Conclusion

In this work, we reconstructed a genome-scale model of *Acinetobacter baylyi* metabolism from the annotation of its genome, metabolic knowledge reported in the literature, and results of high-throughput experiments. The model provides a curated and structured representation of this species's metabolism for use both as a reference and as a foundation for further study. The reconstruction accounts for 875 reactions, 701 distinct metabolites, and 774 genes, and includes nearly all metabolic routes and biochemical conversions identified for *A. baylyi*. A significant proportion of reactions belong to pathways of secondary metabolism that are characteristic of *A. baylyi*'s physiology and lifestyle. The model thus reflects the specific ability of *A. baylyi* to utilize various chemicals originating from plant metabolism, e.g. aromatic acids, hydroxylated aromatic acids, or straight chain dicarboxylic acids. It may assist or even drive future investigations on this bacterium, helping for instance interpret other types of experimental data beyond growth phenotypes, or engineer its metabolism. An increasing number of metabolic engineering strategies are being designed with the help of genome-scale metabolic model predictions [43,44]: the availability of the *A. baylyi* model should facilitate efforts towards biotechnology goals. The *A. baylyi* model may also serve as a basis for the reconstruction of metabolic models of the pathogen strains *Acinetobacter baumanii*. These strains, which are involved in serious nosocomial infections worldwide and have acquired multidrug-resistance capabilities[13], share a significant number of metabolic genes with *A. baylyi* [45]. This model is also the fourth genome-scale bacterial metabolic model to be accompanied by an exhaustive mutant library (with *E. coli* [5,12], *Bacillus subtilis* [4,6], and *Pseudomonas aeruginosa* PAO1 [46,47]). The proximity between *A. baylyi* and *P. aeruginosa*, and to a lesser extent *E. coli*, and the availability of model/mutant library pairs provides an invaluable setup for comparing the metabolism of different species [8].

Several rounds of comparisons of model predictions to large-scale experimental results led to significant model improvements. First, growth phenotypes of the wild-type

strain on 190 distinct environments resulted in the addition of 9 transporters and 2 pathways to the model. After improvement, the model accounted correctly for the growth phenotypes on 173 of the 190 environments. Secondly, we assessed the model against gene essentiality results on 9 defined environments. In contrast with wild-type growth phenotypes, these data can bring indirect information on the gene functions or on the existence of alternate pathways. Investigation on the causes of inconsistencies led us to modify the model in 56 cases out of 124 inconsistent predictions. All model components were modified, the GPR component gathering most of the improvements. The model accuracy in predicting mutant growth phenotypes increased from 88% to 94% on succinate-supplemented minimal medium and from 93% to 94% for the combined conditional gene essentiality results on 8 media. High-throughput phenotype clearly improved the quality of the model and expanded our understanding of *A. baylyi* metabolism, providing a valuable complement to the annotation and the literature. The refinement process was particularly useful in validating or contradicting functional annotations that stood in the "grey zone", i.e. for which the annotation process provided only medium-level evidence.

Conversely, the model allowed systematic evaluation of the results of these high-throughput experiments by comparing them to its predictions. Inconsistencies directly targeted informative experimental results for which further investigation are required. As shown in this work, not all inconsistencies led to model improvements. Some of them could be interpreted in terms of biological processes lying outside the scope of the modeling framework, probably regulation in most cases. In addition, a significant number of discrepancies reported in this work remained unexplained or led to hypotheses in need of confirmation through further study.

The process described here was driven by expert curation: each inconsistency was manually examined in order to search for an interpretation and a possible model correction, a labor-intensive proposition. The systematic use of such experimental data for model refinements would be greatly facilitated by the development of computational methods assisting the curator with his task, however. A number of methods have been developed to search for variants of model which match better with additional experimental data, mainly by seeking additions or removals of reactions in the metabolic network [48,49]. These methods have already proven efficient at suggesting metabolic pathways that account for previously unexplained growth on specific environments [48]. While they can be adapted to handle growth phenotypes of knockout mutant strains, they do not involve the gene-reaction association component of the model, which is shown here to be the main area of model improvement. The association between genes and reactions can be complex as regulatory constraints may interfere with the actual gene function assignments. Computational strategies are therefore needed to help interpret the consequences of gene essentiality data on gene activities.

Deriving the full benefits from a metabolic model entail both accessing its components and using its predictive capabilities. We realized the former by providing access to a detailed metabolic pathways database, the latter through a software tool that performs online predictions, both being coupled at the level of genes and reactions and accessible through a single, highly-interactive interface. This interface allows end-users to carry systems level predictions, and compare them with corresponding experimental observations, putting the consequences of modeling in the context of the detailed biological information that went into the model. This tool should therefore provide researchers interested in *A. baylyi* metabolism with a valuable resource for investigating its phenotypic and physiological properties.

Methods

Initial reconstruction process

The initial reconstruction of the metabolic network was carried out using data provided by (i) the genome expert annotation [19], (ii) the BioCyc metabolic pathway database automatically generated from these annotations [21] and (iii) various literature resources on biochemistry, including textbooks, reviews and journal publications (see Additional file 2). The genome annotation was downloaded from the MaGe interface [50,51] and used as input of the Pathway Tools software [21] in order to generate a BioCyc automatic reconstruction of the metabolic network. The predicted pathways were classified into 7 metabolic categories (central metabolism, nucleotide metabolism, amino acids metabolism, lipid & cell wall metabolism, degradation pathways, cofactor biosynthesis, transport) and examined manually before being included in the model. In order to meet the requirements of the modeling framework the mass balance and reversibility of the reactions were checked.

Reversibility of the reactions was determined from literature evidence when available or based on simple thermodynamic considerations [52]. Proton translocation efficiencies of reactions of the respiratory chain were assumed to be similar to those of *E. coli* [53]. Resulting P/O ratio can range between 0.5 to 2, depending on the types of cytochrome oxidase and NADH dehydrogenase that are used. Reactions using generic compounds (for example a nitrile or a polymer of undetermined length) were instantiated with defined representative metabolites. In this respect, polymeric pathways were expanded into

chains of specific reactions. Large polymeric molecules such as the acyl carrier protein (ACP) or tRNAs were included in the model when they were involved as substrate cofactors of biochemical reactions. Their specific synthesis was not considered in the model. Dependency between reactions and genes were coded by Gene-Protein-Reaction (GPR) Boolean relationships (see below). Using the Cyclone interface to BioCyc [54], we implemented a simple method based on gene homologies between *Escherichia coli* and *Acinetobacter baylyi* to infer enzyme complexes and find AND Boolean associations between genes. Information from the literature was used to close gaps in the metabolic pathways, include pathways specific to *A. baylyi* that were unknown to the metabolic databases, and check the predicted pathways, for instance for the specificity of the cofactors. Physiological information derived from the literature [15,55-59] was used together with genome annotation tools, e.g. TransportDB [60], to add transport reactions in the model. A generic transport reaction was added to the model for each metabolite shown to be utilized by *A. baylyi*. A fixed biomass composition was chosen according to data found in the literature for strains growing on standard media (see Additional file 4). This biomass composition was used to build the reduced list of essential biomass precursors and derive a biomass reaction for Flux Balance Analyses (see below). To help properly account for all metabolic requirements associated with growth, we decomposed the biomass reaction into a set of intermediary biomass reactions synthesizing generic cell constituents (e.g. protein, DNA, RNA, or lipid) from precursor metabolites and a global growth reaction consuming them according to the chosen biomass composition. See Additional file 4 for details on these reactions.

Modeling framework

The metabolic model is composed of three components, namely GPR, NETWORK and BIOMASS. The GPR component models the dependency between genes and reactions using Boolean functions usually called gene-protein-reaction (GPR) associations [22]). For each reaction, a Boolean rule encodes how genes are related to the activity. Genes that are required together are linked with an AND relation while isofunctional genes are linked with an OR relation. The set of GPR associations yields the set of potentially active reactions given the set of available genes.

The NETWORK component models the metabolic network using the constraint-based modeling framework [3]. This framework describes the distributions of reaction fluxes that are compatible with constraints that derive from basic physical assumptions or specific biological information. They are usually formulated as linear constraints, which allow to explore the fluxes solution space using linear programming tools. The main constraint is imposed by the steady-state assumption, represented by the matrix equation:

$$S \cdot v = 0$$

where S is the stoichiometric matrix of the metabolic network and v the vector of reaction fluxes. The stoichiometric matrix is a matrix of size $(m \times n)$ where m is the number of metabolites and n the number of reactions. Each element $S_{i,j}$ of the matrix represents the relative stoichiometric coefficient of metabolite i in reaction j. Additional constraints on the fluxes, such as irreversibility and capacity constraints, are imposed by inequalities in the form:

$$v_{lb,\,i} \leq v_i \leq v_{ub,\,i}$$

where $v_{lb,i}$ and $v_{ub,i}$ are respectively the lower and upper bounds of the flux of reaction i.

Environmental conditions are applied to the model by constraining the exchange fluxes of extracellular metabolites. Exchange fluxes are sink reactions allowing to control the input or output of metabolites in the model. They are constrained to $0 \leq v_i \leq \infty$ for metabolites absent from the medium and $-\infty \leq v_i \leq \infty$ for metabolites present in the medium, except for limiting nutrients for which a maximum uptake rate is chosen ($-v_{uptake} \leq v_i \leq \infty$). When simulating the metabolic network of a knockout mutant, the activity of each reaction is determined by evaluating its GPR association according to the set of removed genes. Fluxes of the inactivated reactions are constrained to be equal to zero.

The BIOMASS component models the essential metabolic requirements for growth. It consists of a list of metabolites that are considered to be essential biomass precursors. Growth phenotype is therefore determined by checking their producibility [26]. To do so, the steady-state constraints for the essential biomass precursors are changed to strict producibility constraints:

$$\begin{cases} S_{internal} \cdot v = 0 \\ S_{biomassprecursors} \cdot v \geq \varepsilon \\ v_{lb,i} \leq v_i \leq v_{ub,i} \end{cases}$$

where $S_{internal}$ is the stoichiometric matrix without the biomass precursors, $S_{biomass\ precursors}$ the stoichiometric matrix restricted to the biomass precursors and ε a vector of small reals, taken as 10^{-3}. Linear programming tools are used to query for a flux distribution fulfilling this set of constraints. If a flux distribution could be found, the model predicted growth, otherwise it predicted no growth.

In order to assess quantitative growth defects, Flux Balance Analyses (FBA) were performed [3]. A biomass reaction was introduced in the model to quantitatively account for the respective contributions of constituent metabolites in the biomass composition (see Additional file 4). Using linear programming, the flux through this reaction was maximized under all constraints, representing the maximal growth rate achievable by the model. Energetic parameters, including growth associated (GAM) and non growth associated (NGAM) maintenance fluxes, were assumed to be similar to those of *E. coli* model [22]. We chose to set NGAM to a constant ATP hydrolysis flux of 10 mmol/h/gDW and GAM to a value of 40 mmol/gDW of ATP in the growth reaction. In all simulations, upper bounds of nutrient exchange fluxes were set to 10 mmol/h/gDW for carbon sources and 100 mmol/h/gDW for other nutrients (see Additional file 2).

Model simulations were performed within FluxAnalyzer [61] and MATLAB® (The MathWorks Inc., Natick, MA) using the YALMIP optimization toolbox [62] and MOSEK optimization solver (Mosek ApS, Copenhagen, Denmark).

Availability of metabolic model

The metabolic model is available both as Excel and SBML files (see Additional files 2 and 5) and will be submitted to the Biomodels.net repository [63]. Whenever possible, cross-references for the model reactions and species to AcinetoCyc [20], KEGG [64] and BiGG [65] databases are provided.

The model is accessible through the NemoStudio web interface [20]. NemoStudio supports growth phenotype predictions, and comparison to experimental results, as well as browsing of model pathways through an interface with AcinetoCyc [20].

Growth phenotyping of the wild-type strain

Growth phenotyping experiments of *A. baylyi* were performed by Biolog, Inc. (Hayward, CA) following experimental procedures described in [66]. Basically, growth of wild-type strains of *A. baylyi* was monitored in PM1 and PM2 microplates containing a defined minimal medium supplemented with 190 distinct carbon sources. The Biolog quantitative growth measures were discretized to yield growth/no-growth qualitative phenotypes by choosing thresholds based on the negative growth control measures and previously known growth phenotypes for *A. baylyi*. Growth phenotypes that were inconsistent with model predictions were checked by examining results from previous work [15], or retesting them individually. Detailed results of Biolog experiments are provided in Additional file 3.

Growth phenotyping of the mutant strains

Detailed experimental protocol for the growth phenotyping of the mutant strains is described in [8]. Basically, using 96-wells plates, the mutant strains were grown in liquid MA minimal media (31 mM Na2HPO4, 25 mM KH2PO4, 18 mM NH4Cl, 41 μM nitrilotriacetic acid, 2 mM MgSO4, 0.45 mM CaCl2, 3 μM FeCl3, 1 μM MnCl2, 1 μM ZnCl2, 0.3 μM (CrCl3, H3BO3, CoCl2, CuCl2, NiCl2, Na2NoO4, Na2SeO3)) supplemented with 25 mM of carbon sources. Succinate/urea medium was composed of MA minimal medium without NH4Cl supplemented with 25 mM of succinate and 20 mM of urea. Absorbance at 600 nm of 24 h cultures was measured to monitor growth. Experiments were performed in duplicates. Measures with discrepant repeats or with weak precultures were discarded from the analyses. Repeats were filtered according to the following rule: a measure was kept if either (1) both repeats were under the growth threshold or (2) the relative difference between the repeats was lower than 50% of the highest value. A threshold of a tenth of the mean absorbance was chosen to classify the mutants in growth or no growth categories. This threshold was chosen particularly low in order to consider as essential only mutants with marked fitness defect.

Authors' contributions

MD reconstructed the initial model, performed model predictions, interpreted inconsistent phenotypes, applied model corrections, and wrote the manuscript. FLF reconstructed the initial model and developed the NemoStudio software tool. VDB participated in the experimental phenotyping and the interpretation of inconsistent phenotypes. AK and DV participated in the initial reconstruction and the interpretation of inconsistent phenotypes. CC and SS developed the NemoStudio software tool. MS participated in the experimental phenotyping and the interpretation of inconsistent phenotypes. JW participated in the design and the coordination of the study. VS conceived of the study, participated in its design and coordination, and contributed to writing the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Sensitivity on GAM and NGAM parameters of growth rate predictions. *This file contains two plots showing the effect of changing growth associated (GAM) and non growth associated (NGAM) maintenance parameters on quantitative growth rate predictions with iAbaylyi*^{v4}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S1.pdf]

Additional file 2

Genome-scale metabolic models*. This file contains the description of all model versions as well as information on reactions, species, biomass precursors, modeled environments and literature references used for the model reconstruction.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S2.xls]

Additional file 3

Experimental data and model refinements*. This file gathers the experimental results used for model refinements, the model predictions, and the corrections/interpretations associated to the inconsistent predictions.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S3.xls]

Additional file 4

Determination of biomass composition of* A. baylyi. This file gathers all information used to reconstruct the biomass assembly reactions in the metabolic model.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S4.xls]

Additional file 5

Genome-scale metabolic model in SBML format*. This file contains the latest model iAbaylyi*^{v4} *in SBML format* http://www.sbml.org.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-S5.xml]

Acknowledgements

We would like to thank Pierre-Yves Bourguignon for comments and insightful discussions on this work. We are also grateful to Georges Cohen, Cécile Fischer, Alain Perret, and Agnès Pinet for their help on the fine points of *A. baylyi* biochemistry. We wish to thank the reviewers for their help in improving the manuscript.

We are grateful for the support of the European Networks of Excellence BIOSAPIENS (contract LSHG-CT-2003-503265) and ENFIN (contract LSHG-CT-2005-518254).

References

1. Joyce AR, Palsson BØ: **The model organism as a system: integrating 'omics' data sets.** *Nat Rev Mol Cell Biol* 2006, **7:**198-210.
2. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation.** *Nat Rev Genet* 2006, **7:**130-141.
3. Price ND, Reed JL, Palsson BO: **Genome-scale models of microbial cells: evaluating the consequences of constraints.** *Nat Rev Microbiol* 2004, **2:**886-897.
4. Oh Y-K, Palsson BO, Park SM, Schilling CH, Mahadevan R: **Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data.** *J Biol Chem* 2007.
5. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.** *Mol Syst Biol* 2006, **2:**2006.0008..
6. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, *et al.*: **Essential Bacillus subtilis genes.** *Proc Natl Acad Sci USA* 2003, **100:**4678-4683.
7. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM: **An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants.** *Proc Natl Acad Sci USA* 2006, **103:**2833-2838.
8. de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, *et al.*: **A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1.** *Mol Syst Biol* 2008, **4:**174.
9. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A: **Essential genes on metabolic maps.** *Curr Opin Biotechnol* 2006, **17:**448-456.
10. Papp B, Pál C, Hurst LD: **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast.** *Nature* 2004, **429:**661-664.
11. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Res* 2004, **14:**1298-1309.
12. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol* 2007, **3:**121.
13. Bergogne-Bérézin E, Towner KJ: **Acinetobacter spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features.** *Clin Microbiol Rev* 1996, **9:**148-165.
14. Metzgar D, Bacher JM, Pezo V, Reader J, Döring V, Schimmel P, Marlière P, de Crécy-Lagard V: **Acinetobacter sp. ADP1: an ideal model organism for genetic analysis and genome engineering.** *Nucleic Acids Res* 2004, **32:**5780-5790.
15. Vaneechoutte M, Young DM, Ornston LN, Baere TD, Nemec A, Reijden TVD, Carr E, Tjernberg I, Dijkshoorn L: **Naturally transformable Acinetobacter sp. strain ADP1 belongs to the newly described species Acinetobacter baylyi.** *Appl Environ Microbiol* 2006, **72:**932-936.
16. Young DM, Parke D, Ornston LN: **Opportunities for genetic investigation afforded by Acinetobacter baylyi, a nutritionally versatile bacterial species that is highly competent for natural transformation.** *Annu Rev Microbiol* 2005, **59:**519-551.
17. Gutnick DL, Bach H: **Potential Application of Acinetobacter in Biotechnology.** In *Acinetobacter Molecular Biology* 1st edition. Gerischer U: Caister Academic Press; 2008:231-264.
18. Abdel-El-Haleem D: **Acinetobacter: environmental and biotechnological applications.** *Afr J Biotechnol* 2003, **2:**71-74.
19. Barbe V, Vallenet D, Fonknechten N, Kreimeyer A, Oztas S, Labarre L, Cruveiller S, Robert C, Duprat S, Wincker P, *et al.*: **Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium.** *Nucleic Acids Res* 2004, **32:**5766-5779.
20. **A. baylyi NemoStudio: Acinetobacter baylyi ADP1 model website** [http://www.genoscope.cns.fr/nemostudio-platform/]
21. Karp PD, Paley S, Romero P: **The Pathway Tools software.** *Bioinformatics* 2002, **18(Suppl 1):**S225-232.
22. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.
23. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic Acids Res* 2005, **33:**D334-337.
24. Abbott BJ, Laskin AI, McCoy CJ: **Effect of growth rate and nutrient limitation on the composition and biomass yield of Acinetobacter calcoaceticus.** *Appl Microbiol* 1974, **28:**58-63.
25. du Preez JC, Lategan PM, Toerien DF: **Influence of the growth rate on the macromolecular composition of A cinetobacter calcoaceticus in carbon-limited chemostat culture.** *FEMS Microbiology Letters* 1984, **23:**71-75.
26. Imielinski M, Belta C, Halasz A, Rubin H: **Investigating metabolite essentiality through genome-scale analysis of Escherichia coli production capabilities.** *Bioinformatics* 2005, **21:**2008-2016.
27. Beste D, Hooper T, Stewart G, Bonde B, Avignone-Rossa C, Bushell M, Wheeler P, Klamt S, Kierzek A, McFadden J: **GSMN-TB: a web-based genome scale network model of Mycobacterium tuberculosis metabolism.** *Genome Biol* 2007, **8:**R89.
28. Lewis JA, Escalante-Semerena JC: **The FAD-dependent tricarballylate dehydrogenase (TcuA) enzyme of Salmonella enterica**

converts tricarballylate into cis-aconitate. *J Bacteriol* 2006, **188:**5479-5486.
29. Lewis JA, Horswill AR, Schwem BE, Escalante-Semerena JC: **The Tricarballylate utilization (tcuRABC) genes of Salmonella enterica serovar Typhimurium LT2.** *J Bacteriol* 2004, **186:**1629-1637.
30. Taylor WH, Juni E: **Pathways for biosynthesis of a bacterial capsular polysaccharide. I. Carbohydrate metabolism and terminal oxidation mechanisms of a capsuleproducing coccus.** *J Bacteriol* 1961, **81:**694-703.
31. Schneider BL, Kiupakis AK, Reitzer LJ: **Arginine catabolism and the arginine succinyltransferase pathway in Escherichia coli.** *J Bacteriol* 1998, **180:**4278-4286.
32. Hove-Jensen B, Harlow KW, King CJ, Switzer RL: **Phosphoribosylpyrophosphate synthetase of Escherichia coli. Properties of the purified enzyme and primary structure of the prs gene.** *J Biol Chem* 1986, **261:**6765-6771.
33. Lawther RP, Wek RC, Lopes JM, Pereira R, Taillon BE, Hatfield GW: **The complete nucleotide sequence of the ilvGMEDA operon of Escherichia coli K-12.** *Nucleic Acids Res* 1987, **15:**2137-2155.
34. Velasco JA, Cansado J, Peña MC, Kawakami T, Laborda J, Notario V: **Cloning of the dihydroxyacid dehydratase-encoding gene (ILV3) from Saccharomyces cerevisiae.** *Gene* 1993, **137:**179-185.
35. Sissler M, Delorme C, Bond J, Ehrlich SD, Renault P, Francklyn C: **An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis.** *Proc Natl Acad Sci USA* 1999, **96:**8985-8990.
36. Fujisaki S, Takahashi I, Hara H, Horiuchi K, Nishino T, Nishimura Y: **Disruption of the structural gene for farnesyl diphosphate synthase in Escherichia coli.** *J Biochem (Tokyo)* 2005, **137:**395-400.
37. Parke SA, Birch GG, MacDougall DB, Stevens DA: **Tastes, structure and solution properties of D-glucono-1,5-lactone.** *Chem Senses* 1997, **22:**53-65.
38. Bryan BA, Linhardt RJ, Daniels L: **Variation in composition and yield of exopolysaccharides produced by Klebsiella sp. strain K32 and Acinetobacter calcoaceticus BD4.** *Appl Environ Microbiol* 1986, **51:**1304-1308.
39. Thorne KJ, Thornley MJ, Glauert AM: **Chemical analysis of the outer membrane and other layers of the cell envelope of Acinetobacter sp.** *J Bacteriol* 1973, **116:**410-417.
40. Kaplan N, Zosim Z, Rosenberg E: **Reconstitution of emulsifying activity of Acinetobacter calcoaceticus BD4 emulsan by using pure polysaccharide and protein.** *Appl Environ Microbiol* 1987, **53:**440-446.
41. Covert MW, Palsson BO: **Constraints-based models: regulation of gene expression reduces the steady-state solution space.** *J Theor Biol* 2003, **221:**309-325.
42. Reitzer LJ: **Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine, and D-alanine.** In *Escherichia coli and Salmonella: cellular and molecular biology Volume 1*. Edited by: Neidhart FC. Washington, D.C.: ASM Press; 1996:391-407.
43. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnol Bioeng* 2003, **84:**647-657.
44. Pharkya P, Burgard AP, Maranas CD: **OptStrain: a computational framework for redesign of microbial production systems.** *Genome Res* 2004, **14:**2367-2376.
45. Vallenet D, Nordmann P, Barbe V, Poirel L, Mangenot S, Bataille E, Dossat C, Gas S, Kreimeyer A, Lenoble P, *et al.*: **Comparative analysis of Acinetobacters: three genomes for three lifestyles.** *PLoS ONE* 2008, **3:**e1805.
46. Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, Ernst S, Will O, Kaul R, Raymond C, Levy R, *et al.*: **Comprehensive transposon mutant library of Pseudomonas aeruginosa.** *Proc Natl Acad Sci USA* 2003, **100:**14339-14344.
47. Oberhardt MA, Puchalka J, Fryer KE, Santos VAPMD, Papin JA: **Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1.** *J Bacteriol* 2008, **190:**2790-2803.
48. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation.** *Proc Natl Acad Sci USA* 2006, **103:**17480-17484.
49. Herrgård MJ, Fong SS, Palsson BØ: **Identification of genome-scale metabolic network models using experimentally measured flux profiles.** *PLoS Comput Biol* 2006, **2:**e72.
50. **MaGe (Magnifying Genomes) – Microbial Genome Annotation System** [http://www.genoscope.cns.fr/agc/mage/]
51. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Médigue C: **MaGe: a microbial genome annotation system supported by synteny results.** *Nucleic Acids Res* 2006, **34:**53-65.
52. Ma H, Zeng A-P: **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.** *Bioinformatics* 2003, **19:**270-277.
53. Gennis RB, Stewart V: **Respiration.** In *Escherichia coli and Salmonella: cellular and molecular biology Volume 1*. Edited by: Neidhardt FC. Washington, D.C.: ASM Press; 1996:217-261.
54. Le Fevre F, Smidtas S, Schachter V: **Cyclone: Java-based querying and computing with Pathway Genome Databases.** *Bioinformatics* 2007.
55. Williams PA, Ray CM: **Catabolism of Aromatic Compounds by Acinetobacter.** In *Acinetobacter Molecular Biology* 1st edition. Gerischer U: Caister Academic Press; 2008:99-117.
56. Eby DM, Beharry ZM, Coulter ED, Kurtz DM, Neidle EL: **Characterization and evolution of anthranilate 1,2-dioxygenase from Acinetobacter sp. strain ADP1.** *J Bacteriol* 2001, **183:**109-118.
57. Jones RM, Collier LS, Neidle EL, Williams PA: **areABC genes determine the catabolism of aryl esters in Acinetobacter sp. Strain ADP1.** *J Bacteriol* 1999, **181:**4568-4575.
58. Jones RM, Pagmantidis V, Williams PA: **sal genes determining the catabolism of salicylate esters are part of a supraoperonic cluster of catabolic genes in Acinetobacter sp. strain ADP1.** *J Bacteriol* 2000, **182:**2018-2025.
59. Ratajczak A, Geissdörfer W, Hillen W: **Alkane hydroxylase from Acinetobacter sp. strain ADP1 is encoded by alkM and belongs to a new family of bacterial integral-membrane hydrocarbon hydroxylases.** *Appl Environ Microbiol* 1998, **64:**1175-1179.
60. Ren Q, Kang KH, Paulsen IT: **TransportDB: a relational database of cellular membrane transport systems.** *Nucleic Acids Res* 2004, **32:**D284-D288.
61. Klamt S, Stelling J, Ginkel M, Gilles ED: **FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps.** *Bioinformatics* 2003, **19:**261-269.
62. Löfberg J: **YALMIP: A Toolbox for Modeling and Optimization in MATLAB.** *Proceedings of the CACSD Conference; Taipei, Taiwan* 2004.
63. **BioModels Database** [http://www.ebi.ac.uk/biomodels/]
64. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]
65. **BiGG Database** [http://bigg.ucsd.edu/]
66. Bochner BR, Gadzinski P, Panomitros E: **Phenotype microarrays for high-throughput phenotypic testing and assay of gene function.** *Genome Res* 2001, **11:**1246-1255.
67. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The Comprehensive Microbial Resource.** *Nucleic Acids Res* 2001, **29:**123-125.

9 Synthèse

9.1 Le modèle confronte efficacement données phénotypiques et connaissance du métabolisme

Comme soulevé dans l'introduction de cette thèse, l'interprétation à l'échelle moléculaire des phénotypes de croissance d'un organisme nécessite de prendre en compte une large variété de processus biologiques. L'utilisation d'un modèle métabolique global permet d'orienter cette interprétation du point de vue du métabolisme. En effet, le fonctionnement de l'ensemble des processus métaboliques y est explicitement modélisé et leur lien à la croissance est pris en compte à l'aide d'une réaction de biomasse ou d'une liste de métabolites précurseurs essentiels à la croissance. La définition de ces dernières regroupe en quelque sorte l'action de tous les autres processus biologiques non modélisés contribuant à la croissance et permet effectivement de relier l'action du métabolisme aux phénotypes. En d'autres termes, l'utilisation conjointe d'un modèle (mécaniste) du métabolisme et d'une réaction de biomasse permet d'étudier isolément le rôle du métabolisme dans l'établissement des phénotypes de croissance.

Le modèle tient compte explicitement de plusieurs composantes du métabolisme : principalement (1) des voies de conversions métaboliques, menant des nutriments aux précurseurs essentiels, (2) des voies de régénération des cofacteurs énergétiques, et (3) des associations entre gènes et réactions, représentant la fonction biochimique des enzymes. Leurs contributions aux phénotypes sont ainsi directement prises en compte dans le modèle. En particulier, l'effet sur les phénotypes d'un changement d'environnement de croissance[62] ou de la délétion d'un gène de fonction métabolique – l'objet des travaux expérimentaux utilisés ici – peuvent être prédit et expliqué par le modèle.

[62] Dans l'hypothèse où la liste des précurseurs essentiels de biomasse ne change pas avec l'environnement. Cette hypothèse, relativement correcte pour des environnements proches (p. ex. les différents milieux minimaux utilisés ici), peut devenir complètement fausse lorsque les environnements exigent des adaptations physiologiques différentes de la cellule. Nous en discuterons plus en détail dans la section sur les limites de l'approche (section 9.4).

Le modèle étant construit à partir de la connaissance actuelle du métabolisme, il prédit les phénotypes attendus d'après cette connaissance. Leur comparaison aux phénotypes observés expérimentalement permet alors de confronter indirectement cette connaissance locale aux observations phénotypiques et d'identifier ainsi des incohérences potentielles entre les deux échelles. Alors que les phénotypes cohérents corroborent d'une certaine manière la connaissance du métabolisme, les phénotypes incohérents y pointent potentiellement des erreurs ou des lacunes. Ils constituent de ce fait un point de départ pour des investigations ultérieures qui permettront de compléter la connaissance du métabolisme. Ce raisonnement peut être effectué « manuellement » pour interpréter les phénotypes ; l'article présentant la banque de mutants d'*A. baylyi* illustre d'ailleurs ce raisonnement pour quelques voies métaboliques précises : biosynthèse de la méthionine, du pantothénate et de l'ubiquinone (de Berardinis et al. 2008). Cependant, l'utilisation du modèle permet de l'automatiser et de confronter ainsi systématiquement tous les phénotypes observés aux phénotypes attendus. De cette manière, tous les phénotypes observés sont évalués en regard du fonctionnement attendu du métabolisme et les phénotypes incohérents sont directement détectés. Avec plus de 4500 phénotypes de croissance[63] à examiner dans le cas d'*A. baylyi*, le gain apporté par la prédiction automatique des phénotypes attendus est notable.

Le travail réalisé sur *A. baylyi* nous a montré que les données phénotypiques complètent utilement celles utilisées pour reconstruire le modèle métabolique initial (annotation du génome et connaissance initiale de quelques voies métaboliques). Ce constat s'applique aussi bien aux phénotypes de croissance de la souche sauvage sur les nombreux environnements Biolog – qui ont contribué à compléter la connaissance des capacités de transport et de catabolisme – qu'aux phénotypes de mutants – qui ont participé à évaluer le fonctionnement des voies métaboliques. Une partie significative des annotations et des voies métaboliques ont en effet été élucidées par analogie avec celles des organismes modèles, principalement *E. coli*. Il est donc probable qu'une partie des activités biochimiques spécifiques à *A. baylyi* n'aient pas été détectées et demeure inconnue. Les phénotypes de croissance apportent justement des données

[63] 190 phénotypes Biolog de la souche sauvage (étape 1 de l'article) + 767 phénotypes de croissance de mutants sur succinate (étape 2) + 8 milieux × 455 mutants = 3640 phénotypes de croissance de mutants sur les 8 milieux minimaux (étape 3).

expérimentales propres à *A. baylyi*, dont l'exploitation est à même d'identifier des incohérences dans cette reconstruction et de guider les corrections et la recherche d'activités propres à *A. baylyi*. De plus, comme évoqué dans l'article, les phénotypes ont également permis de corroborer la fonction métabolique de certains gènes ayant un indice de confiance moyen pour l'annotation.

9.2 Cadre formel d'interprétation des incohérences

Lors de l'étude des phénotypes d'*A. baylyi*, nous avons également montré que le modèle pouvait, en plus d'identifier les phénotypes incohérents, contribuer à rechercher la cause du désaccord. Nous avons pour cela décomposé le modèle en trois composantes représentant chacune des processus biologiques distincts (voir Figure 2A de l'article) :

- GPR, comprenant les règles booléennes d'associations entre gènes et réactions ;
- RESEAU (NETWORK dans l'article), comprenant le réseau de réactions représenté dans le formalisme de la modélisation à base de contrainte ;
- BIOMASSE (BIOMASS dans l'article), comprenant les besoins essentiels de la cellule en précurseurs de biomasse.

Les processus biologiques modélisés dans ces composantes participent tous à la prédiction des phénotypes et peuvent être sources d'incohérence. Leur distinction formelle en trois composantes permet alors de les considérer indépendamment les uns des autres afin d'analyser leur rôle dans la prédiction du phénotype de croissance incohérent.

Au sein de chaque composante, le nombre de types de corrections envisageables à partir d'un phénotype incohérent est limité. Nous en avons regroupé les principaux dans le Tableau 15.

	Prédiction : croissance Observation : non croissance	**Prédiction : non croissance Observation : croissance**
GPR	- une isoenzyme du gène délété n'est pas fonctionnelle - l'isoenzyme supposée du gène délété est en réalité requise simultanément (les protéines forment un complexes au lieu de se remplacer) - le gène délété possède un rôle essentiel supplémentaire	- il existe une isoenzyme au gène délété - le gène délété code pour une sous-unité non-essentielle d'un complexe enzymatique - la réaction associée au gène délété peut s'effectuer spontanément
RESEAU	- une voie métabolique n'est pas fonctionnelle ou n'existe pas	- il existe une voie métabolique supplémentaire (ou alternative à la voie inactivée par la délétion dans le cas d'un mutant)
BIOMASSE	- un précurseur essentiel n'a pas été pris en compte	- un précurseur pris en compte n'est en réalité pas essentiel

Tableau 15. Types de corrections envisageables selon la composante du modèle et la nature de l'incohérence. Ces corrections ne sont bien entendu pas mutuellement exclusives, certaines incohérences pouvant être le fait d'erreurs dans des composantes différentes.

Comme présenté dans l'article, nous avons appliqué ce cadre d'interprétation au modèle d'*A. baylyi* pour rechercher des corrections à ses phénotypes de croissance incohérents. Nous avons systématiquement examiné chacune des incohérences selon ce schéma et recherché pour chaque correction envisagée des indices ou des preuves pouvant la confirmer ou la rejeter. Au total, sur les 124 phénotypes de mutants incohérents, 56 furent corrigés selon ce cadre d'interprétation, concernant pour 34 d'entre elles la composante GPR, pour 12 la composante RESEAU et pour 10 la composante BIOMASSE (voir le Tableau 3 et la Figure 8 de l'article). Des informations supplémentaires sur toutes les incohérences examinées et les corrections retenues sont disponibles dans un fichier Excel à l'adresse : http://www.biomedcentral.com/content/supplementary/1752-0509-2-85-s3.xls .

Ce cadre d'interprétation ne prétend naturellement pas tenir compte de toutes les causes possibles d'incohérence de phénotypes, un grand nombre d'entre elles n'étant pas modélisables. Nous discuterons d'ailleurs plus en détail de ce sujet dans les deux sections suivantes (9.3 et 9.4). Cependant, bien que très simple, il a le mérite d'organiser l'investigation des incohérences de phénotype d'origine métabolique. Comparé aux catégories d'interprétations précédemment proposées par Duarte et al

(2004), ce cadre se limite aux seules causes d'incohérences pouvant être prises en compte dans le modèle mais explicite plus avant les corrections réalisables. De plus, et surtout, il ouvre la voie à une automatisation de la proposition de corrections. En effet, au sein de chaque composante, les types de corrections sont bien définis et déterminés par le sens de l'incohérence. Il est de ce fait envisageable de développer des méthodes proposant de manière systématique des corrections aux composantes du modèle pour les incohérences identifiées. S'agissant de la composante BIOMASSE, Imielinski et al (2005) ont par exemple élaboré des analyses déterminant les métabolites les plus à même d'être essentiels pour expliquer les essentialités de gènes observées. Pour notre part, nous introduirons dans la dernière partie de cette thèse une méthode proposant les corrections d'associations GPR compatibles avec les phénotypes observés. La recherche automatique de corrections dans la composante RESEAU demeure quant à elle, encore plus que pour les autres composantes, un sujet ouvert. Nous donnerons au lecteur des références à des travaux récents dans ce domaine dans la conclusion du manuscrit.

9.3 Exploitation des incohérences non corrigées

Un peu plus de la moitié (68/124) des incohérences détectées sur les phénotypes de mutants d'*A. baylyi* n'ont pas donné lieu à des corrections du modèle. Ces incohérences se répartissent en deux grandes catégories.

Une première partie d'entre elles est constituée d'incohérences dont la cause est déterminée mais qui ne peuvent être corrigées dans le modèle. Les cas causés par des erreurs expérimentales mis à part[64], ces incohérences impliquent généralement des processus biologiques qui ne sont pas modélisés, au premier rang desquels se trouve la régulation. Nous évoquerons ces cas de figures, qui sont associés aux limites du modèle, dans la section suivante, certains d'entre eux pouvant être potentiellement pris en compte en étendant le cadre de modélisation.

La seconde partie de ces incohérences regroupe celles pour lesquelles la cause n'est pas interprétable simplement. Nous avons rencontré 62 incohérences de ce type

[64] Le caractère haut débit et massivement parallèle des expériences de phénotypage rend probable l'occurrence d'erreur de mesure, malgré le soin apporté à leur réalisation.

pour *A. baylyi* ; pour 32 d'entre elles, nous avons pu avancer des hypothèses d'explication, laissant 30 incohérences sans aucune interprétation. Ces cas d'incohérence appellent tous des investigations ultérieures afin de les élucider, pouvant potentiellement mener à des résultats intéressants. Parmi les incohérences de ce type que nous avons détectées se retrouvent notamment toutes celles relevées dans l'analyse « manuelle » de la banque de mutant (de Berardinis et al. 2008). Pour ces dernières, de Berardinis et al (2008) ont proposé des hypothèses d'interprétation nécessitant des recherches supplémentaires pour les valider : principalement la recherche de voies ou d'enzymes alternatives (cas des gènes *panD*, *pdxB*, *ubiC* mentionnés dans cet article) et la démonstration de la non occurrence d'une voie alternative (cas des gènes *pyrC*, *pyrC2*, *metZ* mentionnés dans cet article). Nous avons répertorié dans le fichier Excel mentionné ci-dessus l'ensemble des hypothèses que nous avons formulées pour les incohérences non corrigées. Une partie d'entre elles font d'ailleurs l'objet de recherches dédiées au sein du laboratoire Thesaurus.

La recherche d'une interprétation à une incohérence est souvent plus évidente dans le cas d'un phénotype létal non prédit. En effet, puisque le modèle prédit la croissance de l'organisme, il propose une distribution de flux métaboliques assurant la synthèse tous les précurseurs de biomasse. L'examen de cette distribution révèle ainsi les voies alternatives employées par le modèle pour assurer la croissance ; l'interprétation de l'incohérence consiste alors souvent à rechercher des raisons pour lesquelles ces voies ne seraient en réalité pas actives : existence d'une régulation inhibitrice, enzyme alternative non fonctionnelle par exemple. À l'inverse, l'interprétation d'un phénotype viable non prédit ne bénéficie pas d'une telle assistance du modèle. Quand bien même celui-ci contribue à identifier les métabolites dont la synthèse n'est plus possible, la recherche de nouvelles voies ou d'enzymes alternatives à même de remplacer la fonction inactivée reste un problème ouvert.

En résumé, cette liste des incohérences non corrigées représente l'ensemble des discordances détectées par le modèle entre les phénotypes observés et le réseau métabolique connu. Elle invite donc à des investigations ultérieures afin d'élucider le comportement métabolique réel d'*A. baylyi* expliquant les phénotypes observés et de compléter la connaissance de son métabolisme.

9.4 Limites

9.4.1 Interprétation des phénotypes de croissance faible

L'interprétation de phénotypes de croissance à l'aide d'un modèle métabolique est bien adaptée lorsque les phénotypes observés sont nets, c'est-à-dire distinguables sans ambiguïté entre le cas *létal* et le cas *viable*. Dans la majorité des cas, la létalité nette provoquée par la délétion d'un gène métabolique correspond à l'inactivation complète d'une fonction métabolique, un effet qui est directement pris en compte par le modèle.

Des difficultés apparaissent pour les cas limites, lorsqu'on observe la croissance de la souche mutée, mais à un rythme significativement plus faible que celui de la souche sauvage. Dans notre étude, nous avons appliqué un seuil de croissance relativement bas pour effectuer la distinction qualitative entre mutant viable et non viable (nous l'avons fixé à 1/10 de la croissance de la souche sauvage, voir Matériels et méthodes de l'article). De ce fait, nous avons ainsi considéré comme létales uniquement les délétions provoquant une chute très marquée de la capacité reproductive, d'autant plus que la croissance des mutants fut évaluée de manière clonale et non pas en compétition au sein d'une population hétérogène (voir chapitre introductif, section 2.2.1). Nous avons estimé que cette définition du statut *létal* correspondait au plus proche à la prédiction qualitative de létalité obtenue par le test de productibilité des précurseurs de biomasse (« metabolite producibility », voir Matériels et méthodes de l'article). Cependant, le choix de ce seuil place indistinctement dans la catégorie viable tous les mutants présentant une croissance, même très réduite.

Afin d'aider l'interprétation des cas incohérents pour lesquels la croissance observée est faible, nous avons également déterminé pour chacun des mutants le taux de croissance prédit par la méthode FBA. Cette méthode, qui tient compte des besoins énergétiques de la cellule (voir sections 6.2.8 et 7.2), fournit un résultat quantitatif permettant dans certains cas d'interpréter des diminutions du taux de croissance.

La prédiction de croissance par FBA étant strictement plus contraignante que celle par productibilité des précurseurs[65], tous les phénotypes létaux prédits par test de productibilité le sont également par FBA (taux de croissance nul). La réciproque n'est en revanche pas nécessairement vraie : même si la très grande majorité des phénotypes prédits viables par productibilité des précurseurs le sont aussi par FBA, certains cas présentent des taux de croissance très faibles voire nuls par FBA, contredisant la prédiction initiale (voir Figure 36).

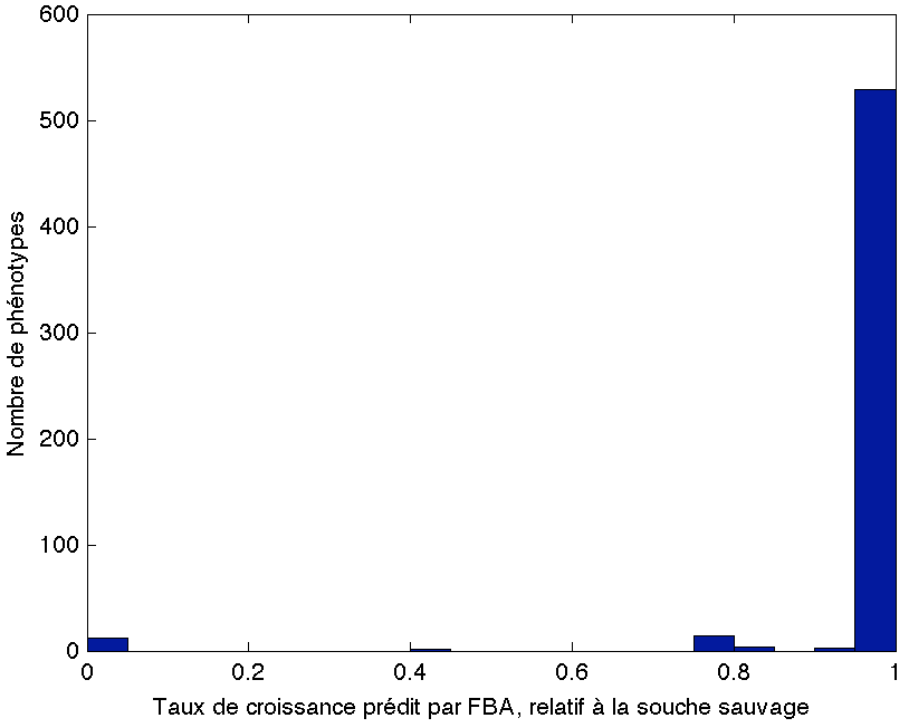


Figure 36. Taux de croissance relatifs à la souche sauvage prédits par Flux Balance Analysis (FBA) pour les mutants prédits viables par analyse de la productibilité des précurseurs de biomasse. Les prédictions ont été effectuées sur milieu minimum avec le succinate comme source de carbone.

Douze mutants de l'étape 2 de raffinement (essentialités sur succinate) sont ainsi prédit viable par test de productibilité mais présentent des taux de croissance prédits par FBA nuls (voir Figure 36). Dix d'entre eux correspondent à des incohérences.

Parmi eux, nous retrouvons neuf gènes essentiels à l'activité de l'ATP synthase (*atpA, atpB, atpC, atpD, atpE, atpF, atpG, atpH, atpI*). L'inactivation de l'ATP synthase n'est en effet pas prédite comme étant létale par le test de productibilité des

[65] La réaction de biomasse utilisée par la méthode FBA contient tous les précurseurs de biomasse utilisés par le méthode de test de productibilité.

précurseurs, ce qui est contredit par leur létalité observée expérimentalement. Pour ces mutants, le modèle réussit à produire tous les précurseurs de biomasse en utilisant l'énergie produite uniquement par phosphorylation au niveau du substrat. Cependant, ce processus ne peut en réalité pas subvenir à tous les besoins énergétiques de la cellule : la méthode FBA, qui tient compte de ces besoins de manière plus complète grâce à la réaction de biomasse, prédit effectivement un taux de croissance nul pour ces mutants.

Le mutant du gène *rpiA* (ribose-5-phosphate épimerase) correspond au dernier phénotype incohérent prédit viable par productibilité des précurseurs et létal par FBA. La différence de prédiction n'est pas due ici à la prise en compte des besoins énergétiques mais à la contrainte de consommation des précurseurs en quantité stœchiométriques imposée par la réaction de biomasse utilisée par le FBA (voir Figure 37).

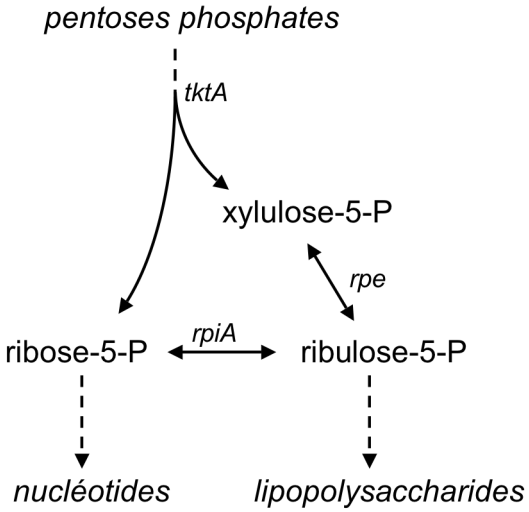


Figure 37. Interprétation de la prédiction du phénotype de croissance du mutant *rpiA*. L'inactivation du gène *rpiA* contraint le ribose-5-phosphate et le ribulose-5-phosphate à être produits en quantité stœchiométriques. Il en est par conséquent de même pour les composés situés à leur aval, notamment des précurseurs de biomasse nucléotides et lipopolysaccharides. La présence d'une contrainte stœchiométrique dans la réaction de biomasse entre ces précurseurs (non compatible avec celle provoqué par la délétion) rend impossible toute croissance du mutant analysée par la méthode FBA. Le test de productibilité des précurseurs ne tient en revanche pas compte de cette contrainte et prédit la croissance du mutant.

La prise en compte des besoins énergétiques de la cellule et la contrainte de consommation stœchiométrique des précurseurs sont les deux seules différences entre ces deux méthodes. L'effet de ces différences sur les prédictions de phénotypes peut tout aussi bien être favorable que défavorable. Comme illustré ci-dessus, prendre en

compte les besoins énergétiques de la cellule peut permettre de prédire la létalité de perturbations des processus énergétiques. Cependant, l'effet d'une telle perturbation dépend également de la capacité de la cellule à s'adapter au manque d'énergie, un paramètre qui n'est pas considéré dans la méthode de prédiction pour laquelle les besoins en énergie restent identique à la souche sauvage. De même, appliquer la contrainte stœchiométrique sur les précurseurs de biomasse peut être aussi bien bénéfique – létalité probable lorsque les déséquilibres de productions de précurseurs sont très marqués – que néfaste – viabilité probable pour de petits déséquilibres. Les deux dernières prédictions divergentes entre FBA et test de productibilité correspondent d'ailleurs à ce dernier cas.

Nous avons choisi de conserver le test de productibilité comme méthode principale de prédiction des phénotypes à la fois pour ces raisons et pour conserver une indépendance entre les précurseurs de biomasse. Cette indépendance facilite en effet les raisonnements de correction de la composante biomasse, dans laquelle seules les présences ou absences des métabolites sont alors à déterminer.

9.4.2 Incohérences d'origine métabolique non prises en compte

Comme évoqué précédemment, une partie des incohérences ayant une cause de nature métabolique ne peuvent être corrigées facilement dans le modèle car ce dernier ne prend pas en compte le processus biologique impliqué. Nous en énumérerons les cas significatifs ci-dessous.

La modélisation à base de contraintes ignore les concentrations de métabolites. Cependant, la perturbation du réseau métabolique peut conduire ces dernières à varier significativement, provoquant potentiellement une accumulation toxique de certains métabolites (Duarte, Herrgard et al. 2004). La prise en compte de cet effet nécessite d'inclure les concentrations métaboliques dans le modèle, ce qui exige alors d'utiliser un cadre de modélisation plus détaillé qui ne peut généralement être mis en œuvre que pour une partie du réseau métabolique global.

Une forte proportion des incohérences qui ne peuvent être corrigées dans les modèles à base de contraintes sont liées à des effets de régulation. Dans un soucis de simplification, ces modèles ignorent en effet complètement les processus contrôlant l'activité des enzymes et supposent que ces dernières sont toutes présentes et actives

en permanence. En réalité, les phénomènes régulateurs contrôlent leurs transcriptions et leurs activités (voir Introduction section 1.2.5) et il est probable que certaines voies alternatives ne puissent suppléer une voie délétée car leurs enzymes ne sont pas suffisamment produites ou sont inhibées. Nous avons par exemple rencontré pour *A. baylyi* le cas de deux enzymes capables de synthétiser indépendamment le glutamate mais ne pouvant se remplacer, chacune d'entre elles étant utilisée sur des plages distinctes de concentrations externes en ions ammonium. Comme mentionné précédemment (voir article de revue, section 3.2.1), les modèles à bases de contraintes peuvent être étendu à l'aide de règles booléennes pour tenir compte des interactions régulatrices agissant sur le métabolisme. Ce formalisme a d'ailleurs été employé avec un certain succès pour interpréter et corriger des phénotypes de croissance incohérents d'*E. coli* (Covert et al. 2004). Cependant, son utilisation nécessite de déterminer les interactions régulatrices à l'œuvre dans l'organisme étudié, une tâche bien plus ardue que la reconstruction du réseau métabolique car ces interactions sont le plus souvent inconnues pour les organismes peu étudiés. L'investigation des phénotypes incohérents peut néanmoins aider à détecter des inhibitions de voies métaboliques et guider ainsi la recherche de ces régulations.

Enfin, l'utilisation d'une composition de biomasse fixe pour prédire la viabilité de l'organisme constitue également une hypothèse restrictive. L'essentialité de certains précurseurs de biomasse dépend en effet des conditions de croissance de l'organisme : un composé non-essentiel dans un environnement peut devenir vital dans un autre. Nous avons par exemple constaté pour *A. baylyi* que certains polysaccharides ne sont pas nécessaires à sa croissance dans les milieux de laboratoire utilisés. Il est néanmoins tout à fait possible que leur production devienne nécessaire dans un environnement plus exigeant, notamment dans des conditions naturelles. Cette hypothèse restrictive est difficile à dépasser car les processus qui rentrent en jeu pour déterminer l'essentialité des précurseurs sont en majorité extérieurs au métabolisme et donc non modélisables dans ce formalisme. Toute méthode capable d'analyser ces processus et de prédire l'essentialité des précurseurs en fonction des conditions extérieures permettrait de dépasser cette limite. À l'inverse, comme illustré dans les travaux d'Imielinski et al (2005), l'étude des phénotypes de mutants peut également aider à déterminer l'essentialité de ces précurseurs à partir de celle des gènes.

10 Extension de l'interface Web de prédiction à d'autres organismes : CycSim

La disponibilité de modèles métaboliques et de phénotypes de croissance de mutants pour un nombre croissant d'organismes nous a conduit à élargir la couverture de l'interface NemoStudio. Nous avons ainsi dérivé de NemoStudio un nouvel outil, CycSim, permettant de contenir les modèles, les bases de données BioCyc et les phénotypes expérimentaux de plusieurs organismes. CycSim a fait l'objet d'une « Application Note » récente dans *Bioinformatics* (Le Fèvre et al. 2009) et est accessible à l'adresse : http://www.genoscope.cns.fr/cycsim . Les fonctionnalités de cette nouvelle interface demeurent quasiment identiques à celles de NemoStudio. Ont été introduites comme nouvelles fonctionnalités :

- des liens directs vers les cartes métaboliques de KEGG,
- l'export des modèles sous la forme de graphes à des fins de visualisation,
- la possibilité d'accéder informatiquement aux modèles à l'aide de services Web,
- la sauvegarde des paramètres d'analyses pour une réutilisation ultérieure.

La version initiale de CycSim contient trois organismes, *Acinetobacter baylyi*, *Escherichia coli* et *Saccharomyces cerevisiae*, basés respectivement sur les modèles iAbaylyi^{v4}, iAF1260 (Feist et al. 2007) et iND750 (Duarte, Herrgard et al. 2004) et sur les bases de données métaboliques AcinetoCyc, EcoCyc (Keseler et al. 2009) et YeastCyc (Christie et al. 2004). Les phénotypes expérimentaux de mutants sont identiques à ceux de NemoStudio pour *A. baylyi* et proviennent (1) dans le cas d'*E. coli* de résultats de phénotypages de la banque de mutant Keio (Baba et al. 2006; Joyce et al. 2006) et de ceux contenus dans la base de données ASAP (Glasner et al. 2003) et (2) dans le cas de *S. cerevisiae* de deux études à grande échelle de phénotypage sur 7 milieux distincts (Giaever et al. 2002; Steinmetz et al. 2002). Cet ensemble de résultats expérimentaux totalise environs 20 000 phénotypes.

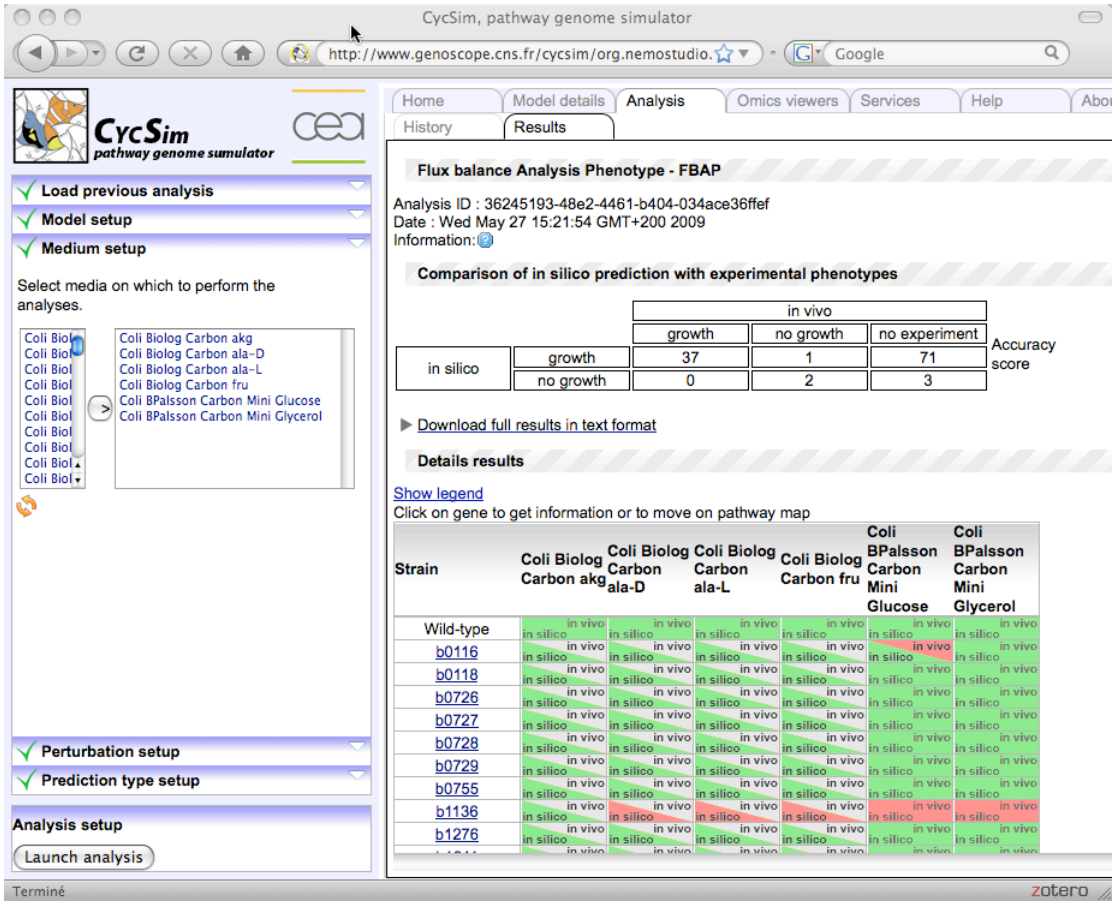


Figure 38. Interface de l'outil de prédiction de phénotypes de croissance CycSim. Code couleur du tableau de résultats : vert, croissance ; rouge, non croissance ; gris, résultat non disponible.

Mes contributions dans ce projet auront été de mettre au point les méthodes de prédiction, d'adapter les modèles extérieurs à nos outils de manière à reproduire correctement les résultats publiés avec ces modèles, de modéliser les environnements de croissance utilisés dans les expériences, et d'effectuer l'interprétation qualitative des phénotypes de croissances quantitatifs.

AUTOMATISATION DE L'INTERPRETATION DES INCOHERENCES D'ORIGINE GENETIQUE

Bénéficiant de l'expérience des corrections appliquées au modèle d'*A. baylyi*, nous avons entrepris de développer une méthode suggérant automatiquement des corrections de la composante GPR des modèles : AutoGPR. Nous consacrerons cette dernière partie sur nos résultats à cette méthode. Dans un premier temps, nous en exposerons le principe et l'implémentation. Nous présenterons ensuite les performances de cette méthode pour retrouver les corrections effectuées au modèle d'*A. baylyi* et l'appliquerons à la résolution des incohérences de trois autres organismes. Enfin, nous discuterons des perspectives d'améliorations et d'applications d'AutoGPR ainsi que de son intégration au sein de stratégies globales de correction des modèles.

11 La méthode AutoGPR

11.1 Principe

L'objectif d'AutoGPR consiste à déterminer automatiquement des modifications aux relations GPR permettant de lever les incohérences de prédictions de phénotypes de croissance. Pour ce faire, la méthode se base fondamentalement sur l'organisation des modèles métaboliques en composantes – GPR, RESEAU et BIOMASSE (voir section 9.2) – afin d'isoler et de manipuler les relations GPR indépendamment des autres composantes du modèle. AutoGPR suppose ainsi les composantes RESEAU et BIOMASSE fixes et correctes et recherche les corrections applicables à la

composante GPR résolvant les prédictions incohérentes de phénotypes sans altérer les prédictions correctes. Cette hypothèse ne préjuge pas de l'existence de corrections dans les composantes RESEAU et BIOMASSE : AutoGPR recherche les corrections GPR compatibles avec une configuration donnée des composantes RESEAU et BIOMASSE. Cette recherche peut être à nouveau effectuée pour toute modification effectuée à ces dernières composantes. Nous discuterons d'ailleurs plus loin dans ce chapitre (section 13.3.2) de la place de la méthode AutoGPR dans une stratégie plus globale de correction des modèles.

Les trois composantes contribuent toutes à prédire les phénotypes de croissance de mutants : la composante GPR prédit les réactions inactivées par la délétion du ou des gènes ciblés tandis que les composantes RESEAU et BIOMASSE prédisent le phénotype de croissance correspondant au réseau métabolique dont ces réactions ont été inactivées (voir Figure 39**A** page 146). Ainsi, du point de vue de la composante GPR, un ensemble de gènes est prédit essentiel si sa délétion inactive un ensemble de réactions dont l'inactivation est prédite létale (par les composantes RESEAU et BIOMASSE) et, inversement, un ensemble de gènes est prédit non-essentiel si sa délétion inactive un ensemble de réactions dont l'inactivation est prédite viable.

Dans le cadre de la prédiction de phénotypes de croissance, les phénotypes d'inactivations de réactions[66] prédits par les composantes RESEAU et BIOMASSE forment l'« interface » entre les GPR et ces deux composantes ; nous les appellerons *essentialité de réaction* par homologie à *essentialité de gène*. AutoGPR supposant les composantes RESEAU et BIOMASSE comme fixes, cette interface l'est également. Ainsi, les essentialités prédites de réactions constituent, tout comme les essentialités observées de gènes, des données extérieures au problème de correction de la composante GPR (voir Figure 39). Toutefois, à la différence des observations expérimentales, ces essentialités de réactions peuvent être connues exhaustivement pour tout ensemble de réactions inactivées en effectuant les prédictions à l'aide du modèle.

[66] En termes plus rigoureux : les phénotypes de croissance prédits pour des modèles dont des ensembles de réactions ont été inactivés.

Nous avons vu précédemment que l'essentialité des gènes dépend de l'environnement dans lequel les phénotypes de croissance ont été observés. Du point de vue du modèle, l'effet de l'environnement est uniquement considéré par la composante RESEAU qui prend en compte l'exploitation des substrats de l'environnement. Les composantes GPR et BIOMASSE, quant à elle, demeurent indépendantes de l'environnement extérieur[67]. Du point de vue de la composante GPR, l'effet d'un changement d'environnement sur les prédictions de phénotypes est ainsi directement pris en compte dans les essentialités de réactions prédites ; à chaque environnement correspondent des essentialités de réactions spécifiques.

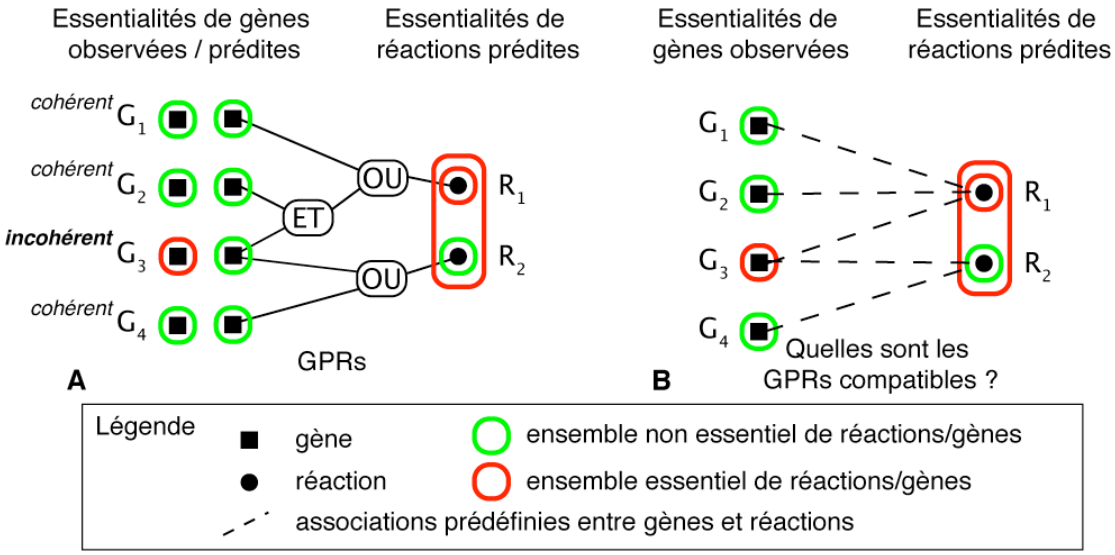


Figure 39. La composante GPR relie l'essentialité des gènes à l'essentialité des réactions. A Les relations booléennes des GPRs combinées aux prédictions d'essentialités des réactions prédisent les essentialités de gènes. Dans l'exemple présenté ici, la prédiction pour le gène G_3 est incohérente avec l'observation expérimentale de son essentialité. Les essentialités de gènes observées et les essentialités de réactions prédites constituent des données extérieures à la composante GPR. **B Principe d'AutoGPR :** générer l'ensemble des relations booléennes respectant les associations prédéfinies entre gènes et réactions et rendant compatibles les essentialités observées de gènes avec les essentialités prédites de réactions.

Plutôt que de chercher à construire un ensemble de corrections-types à appliquer aux GPR existantes (retirer une isozyme, transformer deux isozymes en un complexe, etc.), un processus qui deviendrait rapidement complexe du fait de la composition

[67] pour le cadre de modélisation retenu ici, ce qui n'est pas nécessairement toujours le cas. S'agissant de la composante BIOMASSE, nous avons déjà discuté au chapitre précédent de la pertinence d'utiliser une composition de biomasse dépendant de l'environnement extérieur. D'autre part, des extensions des modèles à base de contraintes pour tenir compte des interactions régulatrices introduisent des relations gènes-réactions dépendant de l'environnement.

possible entre les règles ET et OU, nous avons choisi une approche de type « reverse engineering » visant à construire les GPR à partir des essentialités. La méthode AutoGPR consiste à générer toutes les relations GPR conservant les associations gènes-réactions prédéfinies et rendant compatibles les essentialités prédites de réactions avec les essentialités observées de gènes. Ainsi, sur l'exemple de la Figure 39**B**, AutoGPR génèrera l'ensemble des relations booléennes reliant G_1, G_2 et G_3 à R_1, et G_3 et G_4 à R_2, qui prédisent les essentialités observées de gènes à partir des essentialités prédites de réactions.

Pour chaque réaction, la recherche de relation GPR s'effectue à partir d'un ensemble défini de gènes, délimité par les associations prédéfinies gènes-réactions. Cet ensemble est défini de manière à regrouper tous les gènes susceptibles de participer à la catalyse de la réaction, et donc d'intervenir dans sa GPR. Leur sélection nécessite d'exploiter toute information permettant de les associer à la réaction ; dans la pratique, cette information est contenue dans leurs annotations et la sélection de ces gènes peut être simplement effectué à partir de la GPR initiale de la réaction, qui a elle-même été établie à partir des annotations. La recherche de relations GPR se limite ainsi à cet ensemble prédéterminé de gènes ; les corrections effectuées par AutoGPR n'impliquent de ce fait pas la recherche de nouveaux gènes à associer à la réaction mais plutôt la recherche d'associations booléennes différentes entre gènes déjà identifiés.

La contrainte de rendre compatibles les essentialités de gènes et de réactions à l'aide des GPR peut s'exprimer à l'aide d'une notion d'*impact* de la manière suivante :

- Une délétion de gène(s) est essentielle si et seulement si elle *impacte* (via les GPR) un ensemble essentiel de réaction(s).
- Une délétion de gène(s) est non-essentielle si et seulement si elle *impacte* un ensemble non-essentiel de réaction(s).

Ainsi, chaque essentialité observée de gène contraint les GPR à respecter un comportement défini. En d'autres termes, chaque essentialité observée *spécifie* une partie du comportement des GPR. AutoGPR prend en compte simultanément toutes

ces *spécifications* pour toutes les délétions sur tous les environnements et en déduit l'ensemble des relations GPR compatibles.

Pour résumer, AutoGPR détermine l'ensemble des relations booléennes répondant à ces spécifications et satisfaisant les hypothèses :

- l'ensemble des gènes pouvant être relié à chaque réaction est connu et prédéterminé,
- les relations GPR sont identiques sur tous les environnements,
- les composantes RESEAU et BIOMASSE sont fixes et correctes.

Ces hypothèses permettent à AutoGPR de prendre en compte un grand nombre de types de corrections, dont la recherche d'isoenzymes non fonctionnelles, de sous-unités non-essentielles dans un complexe ou d'associations non identifiées de protéines en complexes. Plus globalement, toute correction impliquant une modification des règles booléennes entre gènes préalablement connus sera identifiée par AutoGPR.

Afin d'expliciter le fonctionnement de la méthode, introduisons des notations mathématiques. Soient :

- **Env** l'ensemble des environnements considérés,
- **R** l'ensemble des réactions du modèle, **P(R)** l'ensemble des sous-ensembles de **R**,
- **G** l'ensemble des gènes du modèle, **P(G)** l'ensemble des sous-ensembles de **G**.

Les essentialités observées de gènes et les essentialités prédites de réactions peuvent être entièrement décrites par deux fonctions **PhenoGene** et **PhenoReac** :

$$\begin{array}{lll} \textbf{PhenoGene}: & \textbf{P(G)} \times \textbf{Env} & \rightarrow \{0,1\} \\ & (\{g_1,...,g_n\},M) & \mapsto \begin{array}{l} 1 \text{ si la délétion de } \{g_1,...,g_n\} \text{ est non essentielle sur M} \\ 0 \text{ si la délétion de } \{g_1,...,g_n\} \text{ est essentielle sur M} \end{array} \end{array}$$

$$\begin{array}{lll} \textbf{PhenoReac}: & \textbf{P(R)} \times \textbf{Env} & \rightarrow \{0,1\} \\ & (\{r_1,...,r_n\},M) & \mapsto \begin{array}{l} 1 \text{ si la délétion de } \{r_1,...,r_n\} \text{ est non essentielle sur M} \\ 0 \text{ si la délétion de } \{r_1,...,r_n\} \text{ est essentielle sur M} \end{array} \end{array}$$

Nous définirons la fonction **Support** pour décrire les associations prédéfinies entre gènes et réactions :

$$\begin{array}{rccl} \textbf{Support}: & \textbf{R} & \rightarrow & \textbf{P(G)} \\ & r & \mapsto & \{g_1,...,g_n\}, \text{ l'ensemble des gènes pouvant être associés à } r \end{array}$$

Les relations entre gènes et réactions sont décrites par l'ensemble de leurs GPR, que nous noterons **GPR**, mais également, de manière équivalente, par une fonction d'impact **Impact** qui détermine les réactions inactivées pour toute délétion de gènes du modèle :

$\mathbf{GPR} = \{\mathbf{GPR_r} \mid r \in \mathbf{R}\}$ avec $\mathbf{GPR_r}: \{0,1\}^{Ng_r} \rightarrow \{0,1\}$ la relation booléenne liant une réaction r à ses Ng_r gènes,

$$\begin{array}{rccl} \textbf{Impact}: & \textbf{P(G)} & \rightarrow & \textbf{P(R)} \\ & \{g_1,...,g_n\} & \mapsto & \{r \in \mathbf{R} \mid \text{r est inactivé par la délétion de } \{g_1,...,g_n\}\} \end{array}$$

La connaissance de la fonction **Impact** est strictement équivalente à celle de **GPR**. En effet, **GPR** peut être défini à partir d'**Impact** :

$$\forall r \in \mathbf{R},\ \forall b \in \{0,1\}^{Ng_r}\ \ \mathbf{GPR_r}(b) = \left| \begin{array}{l} 0 \text{ si } r \in \mathbf{Impact}(\Delta) \\ 1 \text{ sinon} \end{array} \right.$$

$$\text{où } \Delta = \{\text{g} \in \mathbf{Support}(r) \text{ tel que la valeur de } g \text{ dans } b \text{ est } 0\}$$

et inversement, **Impact** peut être définie à partir de **GPR** :

$$\forall \Delta \in \mathbf{P(G)} \quad \mathbf{Impact}(\Delta) = \{r \in \mathbf{R} \mid \mathbf{GPR_r}(b_r) = 0\}$$

$$\text{où } b_r \in \{0,1\}^{Ng_r} \text{ est défini par } b_{ri} = \left| \begin{array}{l} 0 \text{ si le gène } i \text{ de } r \text{ est dans } \Delta \\ 1 \text{ sinon} \end{array} \right.$$

Les contraintes de compatibilité entre les essentialités de gènes et les essentialités de réactions s'expriment désormais simplement par une relation entre les fonctions **PhenoReac**, **PhenoGene** et **Impact** : les relations GPR représentées dans la fonction **Impact** sont compatibles avec les essentialités si et seulement si

$$\mathbf{PhenoGene}(\Delta, M) = \mathbf{PhenoReac}(\mathbf{Impact}(\Delta), M)$$

pour tout couple d'environnement *M* et de délétion de gènes *Δ* pour lequel on dispose d'une observation d'essentialité. Il est à noter que ces observations d'essentialités sont limitées par les résultats expérimentaux d'essentialité de gènes, étant donné que les essentialités de réactions peuvent être connues exhaustivement à l'aide des prédictions des composantes RESEAU et BIOMASSE.

AutoGPR se base sur cette relation et sur la connaissance des valeurs des fonctions **PhenoGene** et **PhenoReac** pour déduire des informations sur la fonction **Impact** avant de déterminer les relations GPR admissibles.

La méthode procède pour cela en deux étapes, une étape de *spécification* suivie d'une étape d'*implémentation* (voir Figure 40).

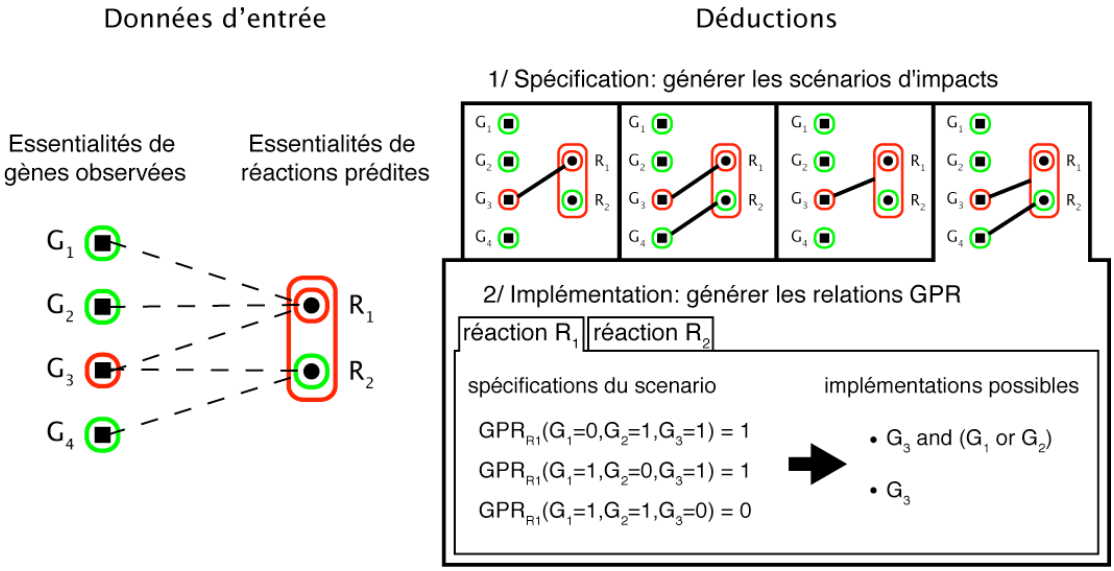


Figure 40. La méthode AutoGPR. AutoGPR déduit l'ensemble des relations GPR compatibles avec les essentialités observées de gènes, les essentialités prédites de réactions et les associations prédéfinies entre gènes et réactions (lignes pointillées). La première étape d'AutoGPR consiste à générer les scénarios d'impacts admissibles compte tenu de ces contraintes : chaque ensemble essentiel de gènes doit impacter un ensemble essentiel de réactions ; chaque ensemble non-essentiel de gènes doit impacter un ensemble non-essentiel (potentiellement l'ensemble vide). De plus, les scénarios d'impacts doivent respecter les associations prédéfinies gènes-réactions. Dans une seconde étape, AutoGPR génère pour chaque scénario d'impacts toutes les relations GPR admissibles. Pour cela, il déduit du scénario d'impacts des spécifications sur les relations booléennes avant de déterminer les implémentations répondant à ces spécifications.

L'étape de *spécification* consiste à envisager l'ensemble des *scénarios d'impacts* compatibles avec les essentialités et les associations prédéfinies gènes-réactions. Pour cela, AutoGPR cherche à attribuer à chaque délétion de gène *Δ* dont l'essentialité est observée une valeur d'impact p_R – un ensemble de réactions inactivées par la délétion – qui :

- respecte les associations prédéfinies gènes-réactions

  $\forall r \in \mathbf{R}, \exists g \in \Delta \text{ tq } g \in \mathbf{Support}(r)$

- rende compatibles sur le milieu M considéré les essentialités des gènes délétés et des réactions inactivées

  $\mathbf{PhenoReac}(p_R, M) = \mathbf{PhenoGene}(\Delta, M)$

Un scénario d'impacts est constitué par de telles attributions réalisées pour chacune des essentialités observées de gènes (voir Figure 40 1/) :

scénario $= \left((\Delta_1, p_{R1}),...,(\Delta_n, p_{Rn})\right)$ pour les n essentialités observées.

Ces scénarios d'impacts étant guidés par les associations prédéfinies gènes-réactions, les valeurs d'impact se limiteront aux ensembles de réactions reliées par ces associations aux gènes délétés. Il est donc suffisant de générer les scénarios d'impacts à l'échelle d'une composante connexe du graphe constitué par les associations prédéfinies et les essentialités de gènes observées pour prendre en compte l'ensemble des cas possibles (voir Figure 41).

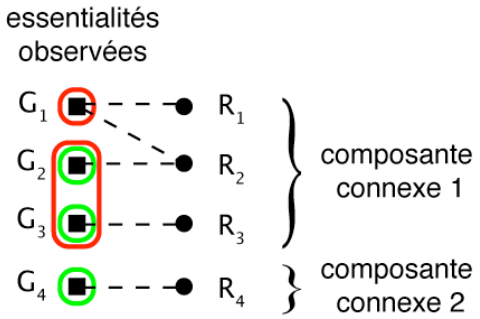


Figure 41. Composantes connexes formées par le graphe des associations prédéfinies gènes-réactions et les groupes de gènes des essentialités observées.

Pour chaque composante connexe incluant une incohérence, AutoGPR génère l'ensemble des scénarios d'impacts envisageables en raisonnant de la manière suivante : si la délétion d'un ensemble de gènes est observée essentielle, celle-ci doit impacter un ensemble essentiel de réactions ; si la délétion est observée non-essentielle, elle doit impacter un ensemble non-essentiel de réactions, possiblement l'ensemble vide. Sur l'exemple de la Figure 40 1/, le dernier scénario d'impact envisagé par AutoGPR est le suivant :

- **Impact**($\{G_1\}$) = ∅
- **Impact**($\{G_2\}$) = ∅
- **Impact**($\{G_3\}$) = $\{R_1, R_2\}$
- **Impact**($\{G_4\}$) = $\{R_2\}$

Si les expériences ont permis de déterminer les essentialités pour plusieurs environnements, seuls les scénarios d'impacts compatibles simultanément avec les essentialités des différents milieux sont alors conservés.

Pour chaque scénario d'impact envisageable, AutoGPR procède ensuite à l'étape d'*implémentation* afin de construire les relations GPR compatibles avec les impacts proposés. Chaque valeur d'impact du scénario d'impact est tout d'abord transformée en spécification sur la relation GPR de chacune des réactions. La Figure 40 2/ détaille par exemple les spécifications déduites sur la relation GPR de la réaction R_1 à partir du dernier scénario d'impact envisagé. Dans un second temps, AutoGPR détermine l'ensemble des relations booléennes entre les gènes et la réaction satisfaisant ces spécifications. Il est important de noter ici que les relations GPR sont des règles booléennes impliquant uniquement les opérateurs ET et OU, la négation n'ayant pas de sens dans une GPR. Cette contrainte, dont AutoGPR tient compte, restreint l'ensemble des relations booléennes implémentables.

Dans certains cas, les contraintes d'essentialités peuvent se révéler incompatibles : aucun scénario d'impact et aucune GPR ne peuvent être générés. AutoGPR en déduit alors qu'une correction purement GPR utilisant les associations prédéfinies gènes-réactions est impossible. Toute correction de l'incohérence dans le modèle implique alors soit de modifier les association prédéfinies gènes-réactions soit de corriger les composantes RESEAU et BIOMASSE pour modifier les essentialités prédites de réactions. Deux cas de figures caractéristiques, dépendant de l'essentialité observée des gènes, sont à l'origine de ces incompatibilités. La Figure 42 les présente sur deux exemples.

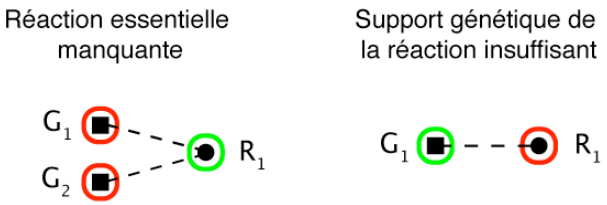


Figure 42. Cas d'incohérences sans correction GPR déductible par AutoGPR. À gauche, les ensembles essentiels de gènes ne peuvent impacter aucun ensemble essentiel de réaction. La correction de l'incohérence nécessite de corriger les composantes RESEAU ou BIOMASSE pour rendre R_1 essentielle ou d'associer G_1 et G_2 à une nouvelle réaction essentielle. À droite, la délétion non-essentielle de G_1 ne peut impacter que l'ensemble vide, impliquant de retirer tous les gènes de la GPR de R_1, ce qui n'est pas admis par AutoGPR. La correction de l'incohérence implique de corriger les composantes RESEAU et BIOMASSE pour rendre R_1 non-essentielle ou d'associer un autre gène à R_1.

Pour résumer globalement le processus, pour chaque composante connexe comportant au moins une incohérence, AutoGPR déduit l'ensemble des scénarios d'impact envisageables à partir des essentialités de gènes observées puis, pour chaque scénario, génère les différentes relations GPR compatibles. Cette méthode propose ainsi de manière exhaustive toutes les configurations possibles de relations GPR permettant de résoudre les incohérences dans la composante connexe considérée.

En pratique AutoGPR peut être utilisé à deux niveaux. Tout d'abord, il permet de déterminer simplement si une correction GPR est envisageable ou non. Si tel n'est pas le cas, le type d'incohérence permet alors de guider la recherche de correction en dehors du périmètre d'AutoGPR (voir Figure 42). Si, en revanche, une correction purement GPR existe, AutoGPR permet ensuite d'énumérer toutes les corrections possibles. Le nombre de relations GPR pouvant devenir extrêmement élevé, des méthodes de sélections des corrections les plus probables peuvent s'avérer nécessaires.

Deux méthodes de sélection sont particulièrement utiles. La première consiste à calculer une distance entre les GPR générées et les GPR initiales du modèle. Cette distance, définie comme le nombre de valeurs distinctes dans les tables de vérité des deux relations booléennes (voir section suivante), cherche à quantifier la différence de « comportement » entre les deux GPR. Nous verrons plus loin dans la partie résultat que les corrections de GPR retenues sont souvent les plus proches des GPR initiales. La seconde méthode de sélection consiste à contraindre des réactions associées aux mêmes gènes à avoir des GPR identiques. Cette simplification est justifiée pour les réactions ayant des activités similaires et opérant sur des substrats très proches (c'est

par exemple le cas des différentes réactions spécifiques dérivées de réactions génériques, voir section 6.2.5). Il est en effet probable dans ces cas que les mêmes enzymes catalysent de manière similaire les différentes réactions. Cette simplification s'applique aisément dans la méthode AutoGPR en remplaçant les réactions similaires par une seule réaction d'essentialité identique à celle de l'ensemble des réactions remplacées, réduisant ainsi le nombre de scénarios d'impacts à considérer (voir Figure 43).

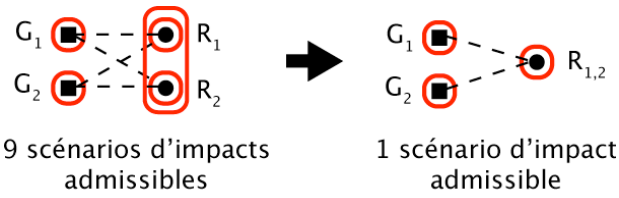


Figure 43. Réduction du nombre de corrections proposées par AutoGPR en imposant des GPR identiques à des réactions. Les réactions R_1 et R_2 sont ici supposées être catalysées de manière identiques par G_1 et G_2. Cette simplification consiste dans AutoGPR à les remplacer par une seule réaction $R_{1,2}$ lors de la génération des scénarios d'impacts.

11.2 Algorithmes

Cette section présente plus en détail les algorithmes utilisés par AutoGPR. Nous l'avons divisé en deux parties. La première détaille les algorithmes développés pour générer toutes les corrections GPR, selon le principe présenté ci-dessus. La seconde introduit une simplification de la méthode permettant de tester plus rapidement la simple existence de correction GPR.

11.2.1 Génération exhaustive des corrections GPR

La génération exhaustive des relations GPR procède comme nous l'avons vu en deux étapes : une étape de spécification et une étape d'implémentation des GPR. Nous présenterons ici les algorithmes utilisés pour réaliser ces étapes.

Calcul des ensembles minimaux de réactions essentielles

La méthode AutoGPR requiert la connaissance de l'essentialité de tous les sous-ensembles de réactions du modèle. Bien que la prédiction d'essentialité d'un ensemble de réactions soit calculable très rapidement (moins de 1/10 de seconde), un calcul exhaustif pour tous les sous-ensembles de réactions se révèlerait beaucoup trop coûteux, le nombre de sous-ensembles augmentant exponentiellement avec le nombre de réactions.

Deux propriétés du modèle et de la méthode AutoGPR permettent heureusement d'en simplifier la tâche.

Tout d'abord, la recherche des scénarios d'impacts est limitée par les associations prédéfinies gènes-réactions. Les ensembles de réactions pouvant être impactés par une délétion de gènes sont donc contenus dans l'ensemble des réactions reliés à ces gènes par les associations prédéfinies. Il est ainsi suffisant de prédire uniquement l'essentialité des sous-ensembles de réactions contenus dans ces ensembles (voir Figure 44). Cette propriété simplifie significativement le calcul des essentialités de réactions, les gènes étant associés majoritairement à un nombre réduit de réactions (voir partie résultat, section 12.1).

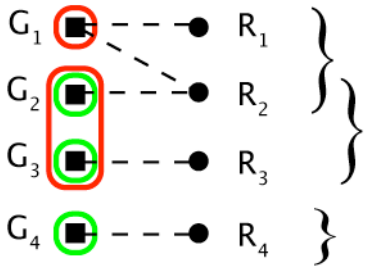


Figure 44. Calcul de l'essentialité des sous-ensembles de réactions. Il est suffisant de limiter le calcul aux sous-ensembles des ensembles de réactions associés aux délétions de gènes. Ces ensembles sont désignés par les accolades.

Ensuite, lorsque les phénotypes de croissance sont prédits par les méthodes FBA ou de productibilité des métabolites, l'essentialité des ensembles de réactions possède la caractéristique d'être monotone par l'inclusion. En effet, dans ces cas, tout ensemble de réactions incluant un sous-ensemble essentiel sera prédit également essentiel. Inversement, tout sous-ensemble d'un ensemble non-essentiel sera prédit non-essentiel.

$$\forall p_{R1}, p_{R2} \in \mathbf{P(R)} \quad p_{R1} \subset p_{R2} \Rightarrow \mathbf{PhenoReac}(p_{R1}) \geq \mathbf{PhenoReac}(p_{R2})$$

Cette propriété découle de la nature de ces deux méthodes de prédictions : elles explorent les distributions de flux admissibles dans le modèle pour déterminer le flux optimal de la réaction de biomasse (FBA) ou vérifier la productibilité de l'ensemble des précurseurs. Un premier modèle ayant un ensemble de réactions inactivées plus grand (par l'inclusion) qu'un deuxième disposera d'un ensemble de distributions de flux contenu dans celui du deuxième. Si aucune distribution de flux du deuxième ne

peut permettre la croissance (flux dans la réaction de biomasse ou productibilité des précurseurs), aucune distribution de flux ne le pourra non plus pour le premier. Inversement, s'il existe une distribution de flux permettant la croissance dans le premier, celle-ci existera également dans le second. Il est à noter que cette propriété de monotonie n'est pas vraie pour toutes les méthodes de prédictions de phénotypes, notamment ROOM ou MoMA (Motter et al. 2008).

Grâce à cette propriété de monotonie, il est suffisant de déterminer les sous-ensembles essentiels minimaux par l'inclusion pour caractériser l'essentialité de tous les sous-ensembles d'un ensemble de réactions. En nous appuyant sur ces propriétés, nous avons utilisé l'algorithme suivant pour déterminer les ensembles essentiels minimaux de réactions :

- Déterminer les ensembles de réactions associés aux délétions de gènes (première propriété, voir Figure 44)
- Pour chacun de ces ensembles
    - Pour chaque environnement
        - Prédire l'essentialité de l'ensemble complet (inactivation conjointe de toutes les réactions)
            - Si non-essentiel : aucun de ses sous-ensembles ne sera essentiel, passer à l'environnement suivant
            - Si essentiel : il contient alors au moins un sous-ensemble essentiel, poursuivre la recherche
        - Pour chaque réaction de l'ensemble
            - Prédire son essentialité
                - Si essentielle : l'ajouter aux ensembles minimaux essentiels pour l'environnement
        - Prédire l'essentialité de l'ensemble des réactions individuellement non-essentielles (inactivation conjointe des réactions prédites non-essentielles)
            - Si non-essentiel : il n'existe pas d'autres sous-ensembles essentiels, passer à l'environnement suivant
            - Si essentiel : il existe des sous-ensembles essentiels impliquant ces réactions, poursuivre la recherche
        - Générer tous les sous-ensembles de l'ensemble des réactions individuellement non-essentielles, ordonnés par la taille
        - Pour chacun de ces sous-ensembles
            - S'il contient un ensemble minimal essentiel déjà identifié : passer au sous-ensemble suivant
            - Sinon : prédire son essentialité
                - Si essentiel : l'ajouter aux ensembles minimaux essentiels pour l'environnement

Le principe de cet algorithme est simple ; il consiste à parcourir tous les sous-ensembles de réactions par ordre croissant de taille et à prédire leur essentialité s'ils

n'incluent pas de sous-ensemble déjà identifié comme essentiel. Il introduit cependant deux tests qui permettent d'épargner de nombreuses recherches inutiles. Il détermine d'une part si l'ensemble de réactions examiné inclut au moins un sous-ensemble minimal essentiel (évitant une recherche inutile dans le cas contraire), et d'autre part s'il existe des sous-ensembles essentiels dans les ensembles de réactions individuellement non-essentielles (teste l'existence de « synthétiques létaux » parmi les réactions, c'est à dire des inactivations multiples létales de réactions individuellement non-essentielles). À l'échelle des ensembles de réactions considérés ici, les réactions « synthétiques létales » se révèlent être relativement rares. Ce second test accélère ainsi significativement la recherche des ensembles minimaux.

Pour des grands ensembles contenant des réactions synthétiques létales, cet algorithme d'exploration systématique peut se révéler trop coûteux. Dans ces cas – que nous avons en pratique très rarement rencontrés pour AutoGPR – des méthodes plus complexes, développées pour déterminer les ensembles essentiels minimaux de gènes, peuvent être employées (Klamt & Gilles 2004; Deutscher et al. 2006; Behre et al. 2007; Imielinski & Belta 2008; Deutscher et al. 2008) (voir également une revue rapide sur ces méthodes en conclusion, section 15).

Génération des scénarios d'impact

Comme évoqué ci-dessus, l'étape de spécification consiste à déterminer exhaustivement tous les scénarios d'impacts qui soient compatibles avec les essentialités et les associations prédéfinies gènes-réactions. Étant donné la nature de ce problème, nous avons choisi d'employer une méthode de programmation logique, implémentée en langage Prolog[68]. La programmation logique se révèle en effet être particulièrement adaptée d'une part à la manipulation d'ensembles et de sous-ensembles d'objets et surtout d'autre part à la recherche par déductions logiques de solutions répondant à des contraintes imposées.

[68] La méthode a été implémentée sous la forme d'un programme Sicstus Prolog http://www.sics.se/sicstus/ .

Le programme Prolog s'appuie sur un ensemble de faits d'entrée définissant :

- Les essentialités observées de gènes
- Les ensembles essentiels minimaux de réactions calculés selon l'algorithme précédent
- Les associations prédéfinies gènes-réactions.

Il définit également des règles permettant d'exprimer la compatibilité d'un impact. Un impact (Δ, p_R) est considéré compatible si les conditions suivantes sont remplies :

- P_R doit être relié à Δ par les associations prédéfinies gènes-réactions
- Si Δ est essentiel, alors p_R doit inclure un ensemble essentiel minimal de réactions.

  Si Δ est non-essentiel, alors p_R ne doit pas inclure d'ensemble essentiel minimal de réactions.

En exploitant conjointement les faits d'entrée et ces règles, le programme Prolog est capable de déduire l'ensemble des impacts compatibles et, à l'échelle d'une composante connexe, de générer les scénarios d'impacts envisageables.

Ces règles définissent la compatibilité des impacts pour un environnement donné. Afin de considérer les contraintes de compatibilité posées par tous les environnements, le programme sélectionne uniquement les scénarios d'impacts valables sur tous les environnements selon le pseudocode suivant :

- Pour chaque composante connexe incluant au moins une incohérence
  - Pour chaque environnement
    - Générer tous les scénarios d'impacts possibles dont les impacts respectent les règles de compatibilité
  - Ne conserver que les scénarios d'impacts existant sur tous les environnements
  - Ecrire les scénarios d'impacts dans le fichier de sortie

Le programme sauvegarde les scénarios d'impacts générés dans un fichier de format XML qui sera utilisé par la suite pour implémenter les GPR.

```
<GPRCORRECT>
<COMPONENT id="Component1" >
<SUPPORT reac="rabaylyi0696" genes="aciad1040 aciad2088 "/>
<SUPPORT reac="rabaylyi0763" genes="aciad0476 aciad2088 "/>
<SCENARIO id="scenario1">
<SPECIFICATION genes="aciad2088 " reacs="rabaylyi0696 rabaylyi0763 "/>
<SPECIFICATION genes="aciad1040 " reacs=""/>
<SPECIFICATION genes="aciad0476 " reacs=""/>
</SCENARIO>
<SCENARIO id="scenario2">
<SPECIFICATION genes="aciad2088 " reacs="rabaylyi0696 rabaylyi0763 "/>
<SPECIFICATION genes="aciad1040 " reacs="rabaylyi0696 "/>
<SPECIFICATION genes="aciad0476 " reacs=""/>
</SCENARIO>
<SCENARIO id="scenario3">
<SPECIFICATION genes="aciad2088 " reacs="rabaylyi0763 "/>
<SPECIFICATION genes="aciad1040 " reacs=""/>
<SPECIFICATION genes="aciad0476 " reacs=""/>
</SCENARIO>
<SCENARIO id="scenario4">
<SPECIFICATION genes="aciad2088 " reacs="rabaylyi0763 "/>
<SPECIFICATION genes="aciad1040 " reacs="rabaylyi0696 "/>
<SPECIFICATION genes="aciad0476 " reacs=""/>
</SCENARIO>
</COMPONENT>
...
```

Figure 45. Format de fichier spécifiant les scénarios d'impacts. Les balises COMPONENT délimitent les informations relatives à une composante connexe. Les balises SUPPORT rappellent les associations prédéfinies gènes-réactions. Les balises SCENARIO définissent chacune un scénario d'impact distinct. Dans un scénario d'impacts, chaque impact est déclaré dans une balise SPECIFICATION.

Implémentation des GPR

L'étape d'implémentation a pour objectif de construire pour chaque réaction les relations GPR compatibles avec les scénarios d'impacts définis à l'étape précédente.

Pour ce faire, nous avons développé un algorithme s'appuyant sur la représentation des fonctions booléennes sous forme de tables de vérités. Partant d'une table de vérité initialement vide, l'algorithme remplit les valeurs correspondant aux spécifications du scénario d'impacts considéré (voir pseudocode ci-dessous et Figure 46). Pour chaque entrée de la fonction booléenne correspondant à une délétion observée de gènes, la valeur spécifiée sera 0 (réaction *inactive*) si la réaction appartient à l'impact de la délétion, et 1 s'il n'appartient pas à l'impact (réaction *active*) (voir Figure 40 page 150 pour un exemple de spécification).

Comme mentionné plus haut, les relations GPR sont des fonctions booléennes utilisant uniquement les opérateurs ET et OU, la négation n'étant pas employée. Cette

caractéristique réduit l'espace des fonctions booléennes à explorer et se traduit par une propriété de monotonie :

$$\forall b_1, b_2 \in \{0,1\}^N \quad b_1 \leq b_2 \Rightarrow GPR_r(b_1) \leq GPR_r(b_2)$$

où la relation d'ordre entre les entrées b_1 et b_2 est incomplète[69].

L'algorithme applique directement la propriété de monotonie aux tables de vérités en complétant les valeurs qui en découlent (voir Figure 46). Il génère ensuite l'ensemble des tables de vérités complètes envisageables, calcule leurs distances à la GPR initiale et les traduits en expressions booléennes sous forme normale disjonctive simplifiée selon le pseudocode suivant :

- Pour chaque composante connexe ayant au moins une incohérence
  - Pour chaque scénario d'impact
    - Pour chaque réaction de la composante connexe
      - Initialiser une table de vérité vide ayant pour entrées les gènes associés à la réaction
      - Remplir la table avec les valeurs spécifiées par les impacts du scénario
        - Si la GPR initiale de la réaction est compatible avec ces spécifications : retourner uniquement cette GPR
        - Sinon : poursuivre
      - Compléter la table de vérité par la propriété de monotonie
      - Si des valeurs de la table de vérité demeurent indéterminées, en déduire toutes les tables de vérités complètes envisageables respectant la propriété de monotonie
      - Pour chaque table de vérité complète
        - Calculer la distance avec la table de vérité de la GPR initiale (nombre de valeurs distinctes)
        - Exprimer formellement la table de vérité en forme normale disjonctive simplifiée[70]
        - Retourner l'expression booléenne obtenue et la distance calculée

[69] $b_1 \leq b_2$ si les éléments de b_1 sont tous un à un inférieurs à ceux de b_2.
[70] À l'aide d'un programme dédié à la manipulation des fonctions booléennes : BDDC v2, disponible à l'adresse http://www-verimag.imag.fr/~raymond/tools/bddc-manual/

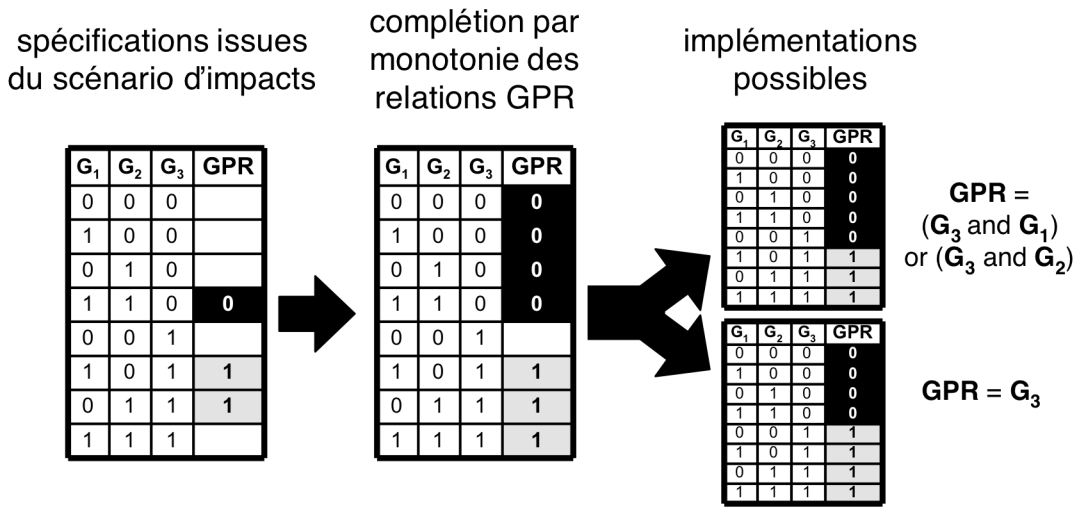


Figure 46. Implémentation de relation GPR à partir des spécifications d'un scénario d'impact.

Nous avons également choisi un format XML pour représenter dans un fichier l'ensemble des GPR générées pour une composante connexe (voir Figure 47). Le nombre de propositions de corrections pouvant devenir très grand, nous avons développé un programme supplémentaire exploitant ces fichiers pour d'une part effectuer des statistiques sur les corrections proposées et d'autre part aider à la sélection des corrections notamment en les ordonnant selon leur distance.

```
<GPRCORRECT>

<COMPONENT id="Component6">
<SUPPORT genes="aciad2983 aciad3598 " reac="rabaylyi0115"/>
<SCENARIO mindist="1" id="scenario1">
<IMPLGPR reac='rabaylyi0115'>
<GPR gpr='aciad2983' dist='1'/>
<GPR gpr='aciad2983.aciad3598' dist='2'/>
</IMPLGPR>
</SCENARIO>

...
```

Figure 47. Format de fichier contenant l'ensemble des GPR générées. Ce format reprend et complète celui utilisé pour énumérer les scénarios d'impact. Pour chaque scénario d'impact, la balise IMPLGPR encadre les implémentations GPR possible d'une réaction. Les balises GPR contiennent chacune une relation GPR compatible, représentée sous la forme d'une expression booléenne (+ représente OU, . représente ET). La distance entre la GPR proposée et la GPR initiale est donnée dans l'attribut dist de la balise GPR.

11.2.2 Test d'existence de correction GPR

La recherche exhaustive de corrections GPR par AutoGPR peut se révéler coûteuse dans les cas où les composantes connexes sont de grandes tailles. Afin d'accélérer le processus de recherche de corrections, nous avons élaboré une méthode simplifiée permettant de tester la simple existence de correction GPR admissible.

Cette méthode fonctionne uniquement pour les cas de délétion simple de gène dont l'essentialité est connue sur un seul milieu. Elle réalise le test d'existence de correction pour chaque essentialité de gène incohérente de la manière suivante.

Si le gène est observé essentiel, une correction GPR existe si et seulement l'ensemble des réactions qui lui sont associées est essentiel. En effet, si tel est le cas, il existe alors au moins un ensemble de réaction qui puisse être impacté de manière compatible par la délétion du gène. Inversement, si tel n'est pas le cas, il n'existe pas d'ensemble de réaction qui puisse être impacté ; aucune correction GPR ne peut être obtenue pour l'incohérence.

Si le gène est observé non-essentiel, une correction GPR existe dès lors que la délétion du gène peut impacter un ensemble non-essentiel de réaction. L'ensemble vide étant non-essentiel, il existerait a priori une solution GPR dans tous les cas, la délétion pouvant être contrainte à n'impacter aucune réaction. Cependant, AutoGPR interdit les corrections aboutissant au retrait de tous les gènes d'une GPR (voir Figure 42 page 153). Le test d'existence d'une correction GPR dans ce cas consiste alors à examiner s'il existe des réactions essentielles reliées uniquement au gène considéré. Si tel est le cas, alors le gène sera contraint à impacter la réaction pour ne pas retirer tous ses gènes, impact incompatible avec l'essentialité de la réaction ; aucune correction GPR n'est de ce fait envisageable. À l'inverse, si tel n'est pas le cas, alors une correction GPR est envisageable dans laquelle aucune des réactions essentielle n'est impactée. Ces dernières étant associées à au moins un autre gène, elles pourront conserver au moins un gène dans leurs GPR, et ce, quel que soit l'impact des autres gènes sur ces réactions[71].

12 Résultats

Afin d'évaluer les performances d'AutoGPR, nous avons appliqué la méthode à la correction des incohérences de cinq modèles métaboliques pour lesquels des données d'essentialités étaient disponibles, parfois sur plusieurs environnements. Toutes les

[71] Par exemple, soit une réaction essentielle reliée à un gène G_1 non-essentiel et à un gène G_2. Si G_2 est non-essentiel, une GPR admissible est « G_1 or G_2 ». Si G_2 est essentiel, une GPR admissible est « G_2 ». Dans tous les cas, la réaction conserve au moins un gène dans sa GPR.

données d'essentialité considérées ici se rapportent à des délétions[72] simples de gènes. Aux deux versions du modèle d'*A. baylyi,* iAbaylyi^{v2} et iAbaylyi^{v3} (considérées séparément car chacune est associée un ensemble de données d'essentialités distinct), nous avons ajouté les modèles de trois organismes supplémentaires, reconstruits par les équipes de deux laboratoires distincts. Le Tableau 16 ci-dessous présente ces différents modèles et les ensembles de données d'essentialité utilisés. Dans un soucis de clarté, nous désignerons ces modèles par le nom de l'organisme en précisant pour *A. baylyi* le numéro de version.

Organisme	Modèle	Référence modèle	Milieux des tests d'essentialité	Références données d'essentialités
A. baylyi	iAbaylyi^{v2}	(Durot et al. 2008)	succinate	(de Berardinis et al. 2008)
A. baylyi	iAbaylyi^{v3}	(Durot et al. 2008)	acetate asparagine butanediol glucarate glucose lactate quinate urea	(Durot et al. 2008; de Berardinis et al. 2008)
E. coli	iAF1260	(Feist et al. 2007)	glucose [1] glycerol [2]	[1] (Baba et al. 2006) [2] (Joyce et al. 2006)
B. subtilis [73]	iYO844	(Oh et al. 2007)	rich (riche)	(Kobayashi et al. 2003)
S. cerevisiae	iND750	(Duarte, Herrgard et al. 2004)	ypd_ess (riche), [3] mmd [3] ypd (riche), [4] ypdge (riche), [4] ype (riche), [4] ypg (riche), [4] ypgal (riche), [3] ypl (riche), [4]	[3] (Giaever et al. 2002) [4] (Steinmetz et al. 2002)

Tableau 16. Modèles métaboliques et données d'essentialités utilisés pour évaluer la performance d'AutoGPR. Les milieux des tests d'essentialité sont minimaux sauf autrement indiqué par (riche). Milieux pour *S. cerevisiae* : mmd, minimal glucose ; ypd_ess, Yeast extract Peptone (YP) + glucose (dataset contenant uniquement les gènes essentiels) ; ypd, YP + glucose ; ypdge, YP + glucose, glycerol et ethanol ; ype, YP + ethanol ; ypg, YP + glycerol ; ypgal, YP + galactose ; ypl, YP + lactate. Les auteurs des modèles reconstruits ayant tous exploité les données d'essentialité pour évaluer leurs modèles, nous avons extrait les données d'essentialités des papiers « modèles ».

[72] Les gènes de *B. subtilis* n'ont en toute rigueur pas été délétés, mais inactivés par insertion d'une cassette de disruption (voir section 2.2.1).

[73] Un nouveau modèle de *B. subtilis,* qui exploite de manière très complète des données d'essentialité, a été publié récemment (Henry et al. 2009). Ce travail est cependant trop récent pour pouvoir être inclus dans nos travaux.

12.1 Complexité des GPR dans les modèles métaboliques

Lors de la correction manuelle du modèle d'*A. baylyi* à partir des incohérences de phénotypes, nous avons constaté qu'une partie significative d'entre elles impliquaient la composante GPR du modèle, d'où notre motivation de développer la méthode AutoGPR pour assister le curateur dans ces corrections.

Cependant, dans le but d'évaluer la pertinence et l'intérêt pratique réel d'une telle méthode de raisonnement automatique sur les GPR, nous avons tout d'abord cherché à obtenir un aperçu de la complexité des GPR dans les modèles considérés ici. Si, en moyenne, les GPR comportaient très peu de gènes et impliquaient peu de relations booléennes distinctes, l'intérêt pratique d'une telle méthode s'avèrerait réduit. Au contraire, si les GPR mettaient en jeu des combinaisons complexes de plusieurs gènes, son intérêt serait a priori plus significatif.

Dans un premier temps, nous avons évalué le nombre de GPR présentes dans ces modèles ainsi que leur variabilité. Le Tableau 17 donne un aperçu global de la taille des modèles et du nombre de GPR distinctes impliquées dans chacun d'entre eux. Dans cette partie, seul le modèle A. baylyi v2 sera considéré, les résultats pour A. baylyi v3 étant quasiment identiques.

Modèle	Nombre de gènes	Nombre de réactions	Nombre de GPR distinctes (% p. r. aux réactions)
A. baylyi	789	993	532 (54%)
E. coli	1260	2382	960 (40%)
B. subtilis	844	1250	586 (47%)
S. cerevisiae	750	1267	546 (43%)

Tableau 17. Nombre de réactions et de GPR distinctes dans les quatre modèles.

Les modèles d'*A. baylyi*, de *B. subtilis* et de *S. cerevisiae* sont de tailles relativement équivalentes. Le modèle d'*A. baylyi* compte moins de réactions mais intègre un nombre équivalent de gènes et de GPR distinctes, suggérant qu'un nombre comparable de processus biochimiques distincts sont pris en compte dans ces trois modèles. Le modèle d'*E. coli* comprend quant à lui significativement plus (quasiment le double) de gènes, réactions et GPR distinctes. Ce résultat traduit d'une part le fait que la connaissance du métabolisme d'*E. coli* est bien plus complète que pour *A. baylyi, B. subtilis* et, dans une moindre mesure, *S. cerevisiae* et d'autre part l'effort de reconstruction plus conséquent pour cet organisme, qui profite à la fois de versions

précédentes du modèle (Edwards & Palsson 2000; Reed et al. 2003) et de la base de données métabolique très complète Ecocyc (Keseler et al. 2009).

Le nombre de GPR distinctes ramené au nombre total de réactions est relativement similaire pour tous les modèles (entre 40 et 54%). Ce ratio s'explique par la présence dans les modèles d'un nombre important de réactions non associées à un gène – notamment des réactions spécifiques à la modélisation, telles que les réactions d'échanges ou d'assemblage de la biomasse – mais également par le fait que certaines GPR sont partagées par plusieurs réactions. La Figure 48 illustre cet effet en traçant les distributions du nombre de réactions associées à chaque GPR distincte pour les quatre modèles considérés ici.

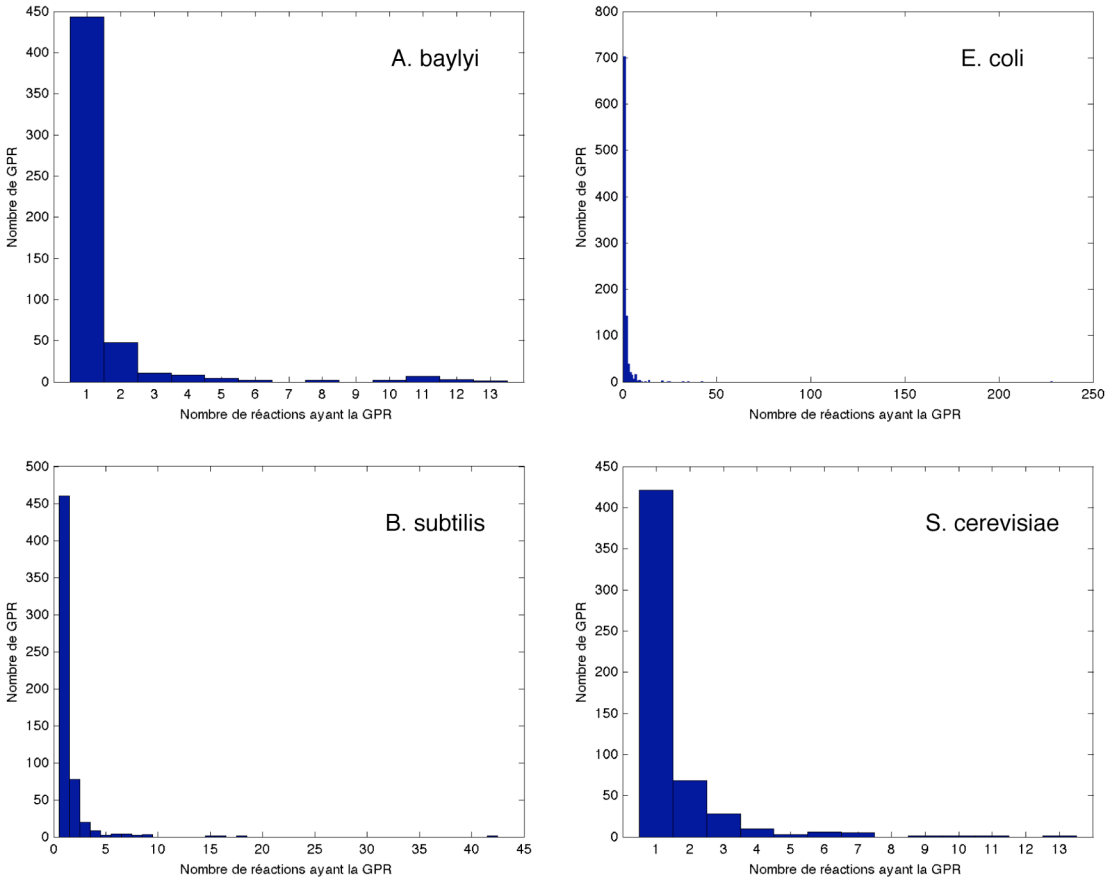
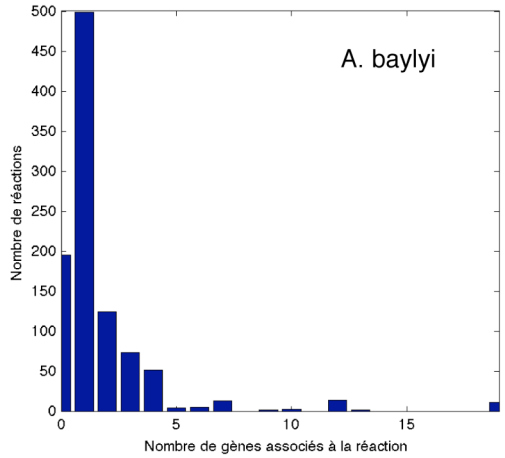


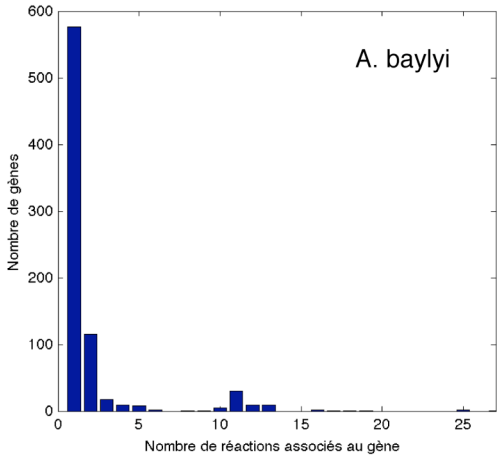
Figure 48. Distribution du nombre de réactions associées à chaque GPR distincte. Les réactions de GPR vide (non associées à un gène) ont été écartées de cette analyse.

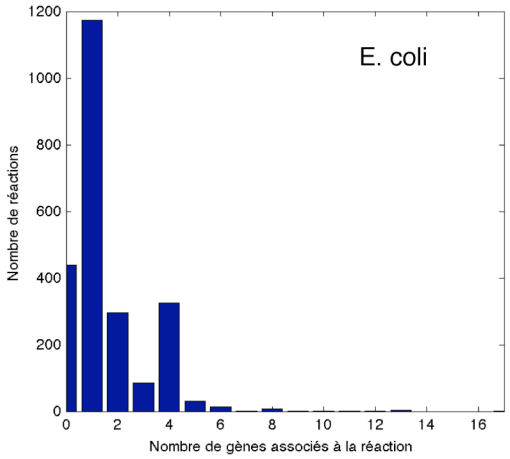
Bien que la grande majorité des GPR soient, dans tous les modèles, associées à une unique réaction, une part non négligeable d'entre elles sont associées à 2 réactions ou plus. Dans ces derniers cas, les GPR représentent des activités biochimiques capables de transformer plusieurs substrats différents. Par exemple,

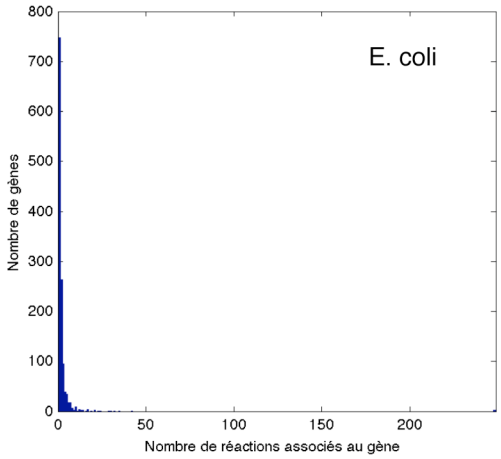
dans chacun des modèles, les GPR partagées par 5 à 8 réactions sont en majorité liées aux processus de synthèse et de dégradation des lipides, pour lesquelles plusieurs réactions agissant similairement sur des lipides de longueurs différentes possèdent la même GPR. La Figure 48 révèle également que certaines GPR possèdent une « spécificité » très large, une GPR d'*E. coli* codant pour une porine est par exemple associée à 228 réactions qui réalisent le transport d'une grande variété de métabolites entre le milieu extéruieur et le périplasme. En résumé, ces premiers résultats nous montrent que le nombre de GPR distinctes dans les modèles est élevé, malgré le fait que certaines d'entre elles soient partagées par de nombreuses réactions.

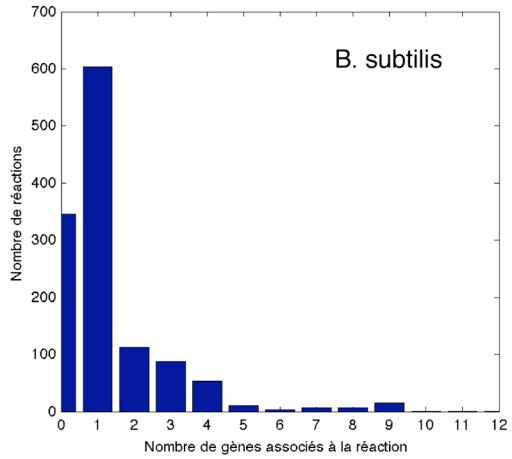
La Figure 49 explore plus avant les interdépendances entre gènes et réactions en traçant les distributions du nombre de gènes par réaction et de réactions par gène.

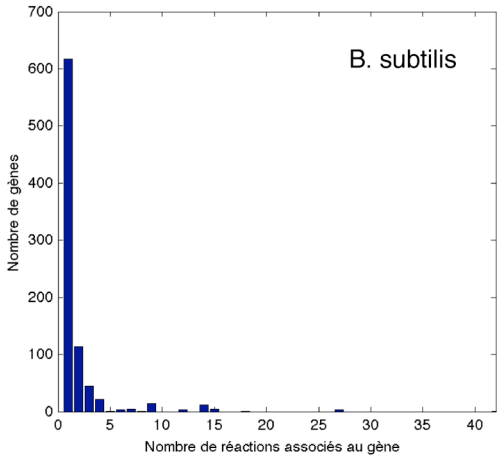

A. baylyi
Nombre de réactions
Nombre de gènes associés à la réaction

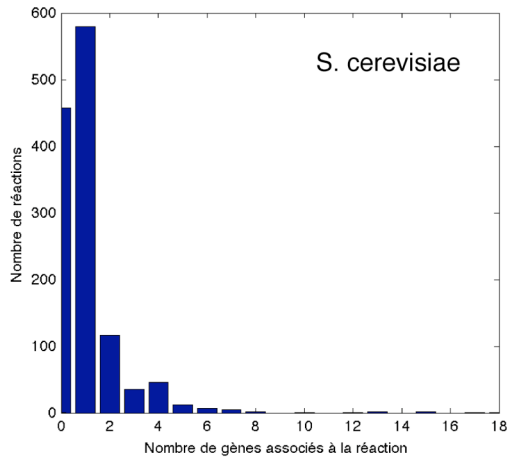

A. baylyi
Nombre de gènes
Nombre de réactions associés au gène


E. coli
Nombre de réactions
Nombre de gènes associés à la réaction


E. coli
Nombre de gènes
Nombre de réactions associés au gène


B. subtilis
Nombre de réactions
Nombre de gènes associés à la réaction


B. subtilis
Nombre de gènes
Nombre de réactions associés au gène


S. cerevisiae
Nombre de réactions
Nombre de gènes associés à la réaction

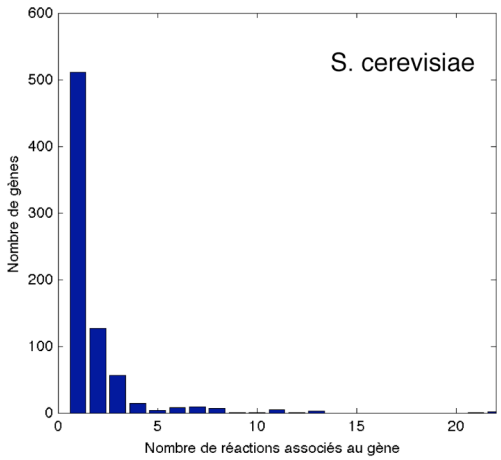

S. cerevisiae
Nombre de gènes
Nombre de réactions associés au gène

Figure 49. Distributions du nombre de gènes associés à chaque réaction et du nombre de réactions associées à chaque gène dans les GPR des quatre modèles.

Les distributions du nombre de réactions par gène illustrent les mêmes tendances que la Figure 48 : un grand nombre de gènes ont une spécificité très faible (liés à peu de réactions) tandis qu'un groupe plus réduit possède une spécificité large.

À l'inverse, les distributions du nombre de gènes par réaction donnent un aperçu du nombre de gènes impliqués dans la catalyse de chaque réaction, et ainsi du nombre de gènes composant les GPR des modèles. Cette statistique est importante car AutoGPR ne peut proposer des corrections que pour les réactions liées à deux gènes ou plus. Les histogrammes de la Figure 49 (colonne de gauche) montrent qu' une part importante des réactions sont effectivement associées à deux gènes ou plus et que, parmi ces réactions, une majorité est même reliée à trois gènes ou plus, rendant les combinaisons de corrections GPR plus complexes et le recours à AutoGPR plus intéressant.

Enfin, dans le but d'estimer la variété des règles booléennes utilisées dans les GPR, nous avons calculé les distributions bivariées du nombre de ET et du nombre de OU dans chaque GPR. Pour cela, et afin de pouvoir comparer rigoureusement les résultats entre modèles, chaque GPR a été exprimée en forme normale disjonctive. La Figure 50 présente ces résultats.

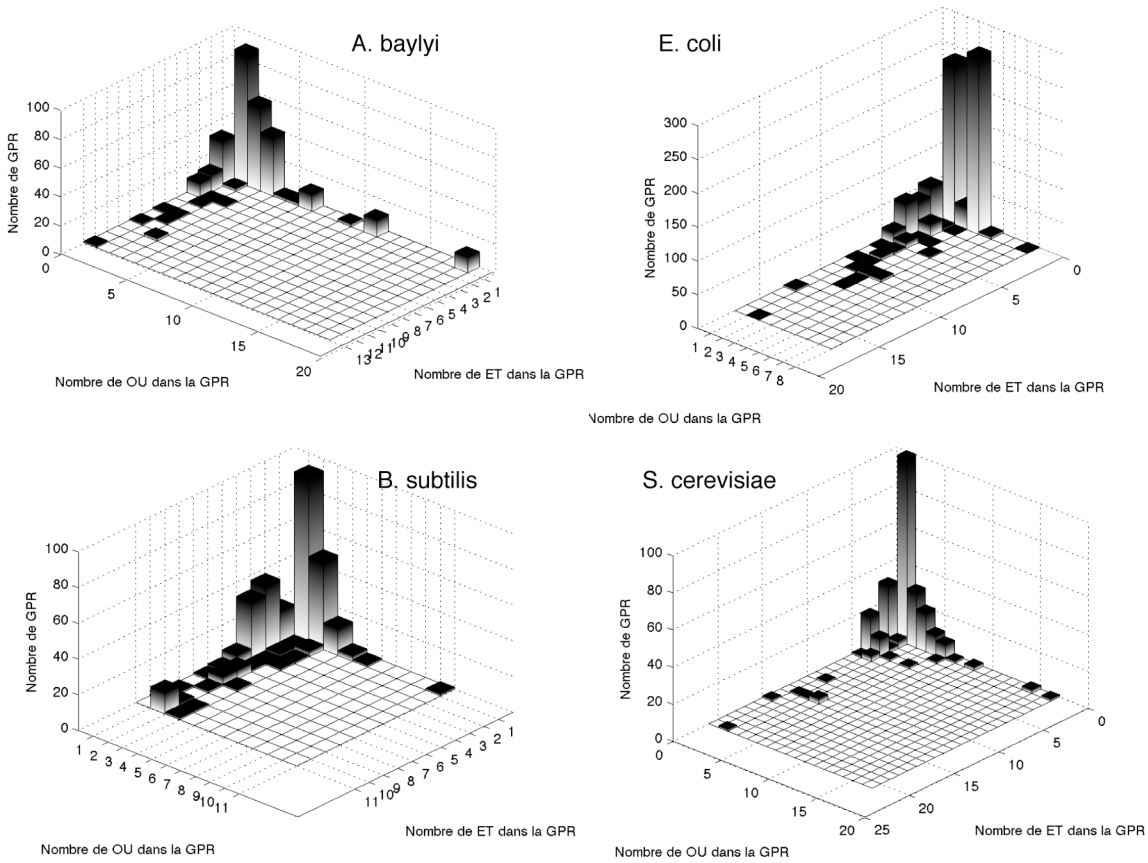


Figure 50. Distribution du nombre de ET et OU dans les GPR des quatre modèles. Pour une comparaison rigoureuse, les relations GPR ont toutes été exprimées en formes normales disjonctives. Par soucis de clarté, le nombre de GPR ne comptant aucun ET et aucun OU n'est pas affiché (ce nombre peut être déduit de la Figure 49, il s'agit du nombre de réactions associées à 0 ou 1 gène).

Pour les quatre modèles, une majorité de GPR possède exclusivement des opérateurs ET ou des opérateurs OU ; ces GPR expriment la présence soit d'un complexe soit d'isozymes. Le nombre d'isozymes et, dans une moindre mesure, de sous-unités de complexe peut être relativement élevé. En effet, les modèles d'*A. baylyi*, d'*E. coli* et de *S. cerevisiae* comptent par exemple un grand nombre de GPR ayant 4 isozymes ou plus. De même, les tailles de complexes dépassent fréquemment 4 pour tous les modèles. Dans chaque modèle, une minorité tout de même non négligeable de GPR inclut simultanément des opérateurs ET et OU, exprimant des alternatives entre complexes ou entre sous-unités d'un même complexe. Bien que relativement peu nombreuses, ces dernières GPR sont cependant susceptibles de comporter des erreurs dont l'interprétation à la lumière des phénotypes de croissance peut se révéler plus complexe.

En conclusion, cette petite étude sur les GPR des quatre modèles nous a montré qu'un nombre significatif d'entre elles ne sont pas triviales et peuvent typiquement bénéficier de la capacité de raisonnement automatique d'AutoGPR.

12.2 Statistiques globales sur les propositions d'AutoGPR

Dans cette partie, nous appliquerons la méthode AutoGPR à chaque modèle pour résoudre leurs incohérences par rapport aux données expérimentales d'essentialité. Nous présenterons ici (1) des statistiques sur les incohérences de chaque modèle, (2) les résultats des tests simples d'existence de correction GPR et (3) les résultats des propositions de correction GPR par AutoGPR.

12.2.1 Confrontation des modèles aux données d'essentialité

Les auteurs des cinq modèles ont tous confronté les prédictions de leurs modèles aux données d'essentialités considérées ici. Afin d'exploiter par la suite leurs interprétations expertes des incohérences, nous avons réalisé les prédictions en utilisant les mêmes méthodes et paramètres, lorsque cela était possible.

Les prédictions de phénotypes avaient toutes été effectuées par la méthode FBA, sauf pour les modèles A. baylyi (test de productibilité des précurseurs). Afin d'uniformiser notre processus de test, nous avons tout de même réalisé les prédictions de tous les modèles par FBA. Comme déjà discuté à la section 9.4.1, ce changement de méthode provoque un petit nombre de différences de prédiction pour *A. baylyi* ; ces différences n'impactent cependant pas significativement l'analyse effectuée ici.

Pour les modèles A. baylyi v2, A. baylyi v3, E. coli et B. subtilis, les contraintes sur les flux d'échanges (modélisant les milieux) et les seuils de croissance à appliquer aux réactions de biomasse (pour déterminer le phénotype qualitatif croissance/non croissance) étaient explicités par leurs auteurs. Nous avons pu reproduire la totalité de leurs prédictions. Dans le cas de S. cerevisiae, ces paramètres n'étaient pas précisés. Nous les avons inférés de manière à reproduire au mieux les résultats des auteurs. Deux ou trois prédictions par milieu demeurent différentes de celles publiées, elles sont dans leur très grande majorité identiques.

Le Tableau 18 présente les prédictions obtenues et leur cohérence par rapport aux données expérimentales.

	Gène observé essentiel			Gène observé non essentiel			Tout gène		
	coh.	incoh.	% coh.	coh.	incoh.	% coh.	coh.	incoh.	% coh.
A. baylyi v2									
succinate	197	54	*78%*	487	29	*94%*	684	83	*89%*
A. baylyi v3									
acetate	2	2	*50%*	423	9	*98%*	425	11	*97%*
asparagine	1	2	*33%*	437	8	*98%*	438	10	*98%*
butanediol	6	4	*60%*	422	11	*97%*	428	15	*97%*
glucarate	5	0	*100%*	406	7	*98%*	411	7	*98%*
glucose	3	4	*43%*	422	10	*98%*	425	14	*97%*
lactate	0	2	*0%*	434	11	*98%*	434	13	*97%*
quinate	4	4	*50%*	426	10	*98%*	430	14	*97%*
urea	3	0	*100%*	435	7	*98%*	438	7	*98%*
tous milieux	17	18	*49%*	403	17	*96%*	420	35	*92%*
E. coli									
glucose	159	78	*67%*	982	29	*97%*	1141	107	*91%*
glycerol	156	85	*65%*	978	29	*97%*	1134	114	*91%*
tous milieux	151	106	*59%*	967	24	*98%*	1118	130	*90%*
B. subtilis									
rich	63	28	*69%*	657	18	*97%*	720	46	*94%*
S. cerevisiae									
ypd_ess	31	87	*26%*	0	0	-	31	87	*26%*
mmd	35	11	*76%*	502	16	*97%*	537	27	*95%*
ypd	3	73	*4%*	476	13	*97%*	479	86	*85%*
ypgal	7	2	*78%*	543	12	*98%*	550	14	*98%*
ypdge	17	64	*21%*	465	19	*96%*	482	83	*85%*
ypg	23	62	*27%*	466	14	*97%*	489	76	*87%*
ype	23	60	*28%*	461	21	*96%*	484	81	*86%*
ypl	22	61	*27%*	466	16	*97%*	488	77	*86%*
tous milieux	65	210	*24%*	379	28	*93%*	444	238	*65%*

Tableau 18. Confrontation des essentialités prédites aux essentialités observées expérimentalement pour les 5 modèles et les différents milieux considérés. Toutes les prédictions d'essentialités ont été réalisées par la méthode FBA. « coh. » et « incoh. » désignent les gènes prédits de manière respectivement cohérente et incohérente par rapport à l'observation expérimentale. La ligne *tous milieux* considère tous les milieux simultanément : un gène est essentiel sur *tous milieux* s'il l'est sur au moins l'un d'entre eux ; il est non-essentiel s'il l'est sur tous les milieux.

Sur l'ensemble des modèles et des milieux, le taux de bonnes prédictions est en moyenne largement meilleur pour les gènes non-essentiels (toujours supérieur à 94%) que pour les gènes essentiels. Nous avions déjà évoqué cette tendance pour *A. baylyi*. Elle s'interprète par le fait qu'une large part des réactions des modèles ne participe pas au fonctionnement métabolique sur des milieux précis ; la fonction des gènes qui leurs sont associés ne peut être réellement évaluée par leur essentialité.

Les taux de bonnes prédictions des modèles A. baylyi v2, A. baylyi v3, E. coli et B. subtilis sont du même ordre de grandeur (sans tenir compte des fluctuations pour A. baylyi v3 pour les gènes essentiels, dues à leur petit nombre). En revanche, les

prédictions de S. cerevisiae pour les gènes essentiels sont significativement moins bonnes. La technique expérimentale utilisée pour déterminer l'essentialité des gènes en est probablement la cause. Alors que pour A. baylyi, E. coli et B. subtilis la croissance des mutants fut évaluée de manière clonale (chaque mutant individuellement, voir section 2.2.1), celle des mutants de S. cerevisiae fut évaluée en compétition au sein d'une population regroupant mutants et souche sauvage (Giaever et al. 2002; Steinmetz et al. 2002). Cette dernière méthode détecte rapidement tout effet d'une délétion sur la capacité reproductive (voir discussion dans l'introduction, section 2.2.2), mais tend à classer essentiels des gènes dont la délétion ralentit simplement la croissance. Les modèles ayant plus de difficulté à prédire une diminution quantitative de croissance qu'une létalité totale, le taux de bonnes prédictions pour les gènes essentiels de S. cerevisiae s'en trouve affecté.

Au total, les nombres d'incohérences à traiter par AutoGPR s'élevent respectivement à 83, 35, 130, 46 et 238 gènes pour A. baylyi v2, A. baylyi v3, E. coli, B. subtilis et S. cerevisiae.

12.2.2 Tests simples d'existence de correction GPR

Dans un premier temps, nous avons testé l'existence de correction GPR. Nous avons utilisé pour cela le test simple d'existence (voir 11.2.2) appliqué séparément aux incohérences de chaque milieu. Ce test n'est en effet valable que pour des délétions simples évaluées sur un milieu unique.

Le Tableau 19 présente les résultats de ces tests appliqués aux incohérences de tous les modèles.

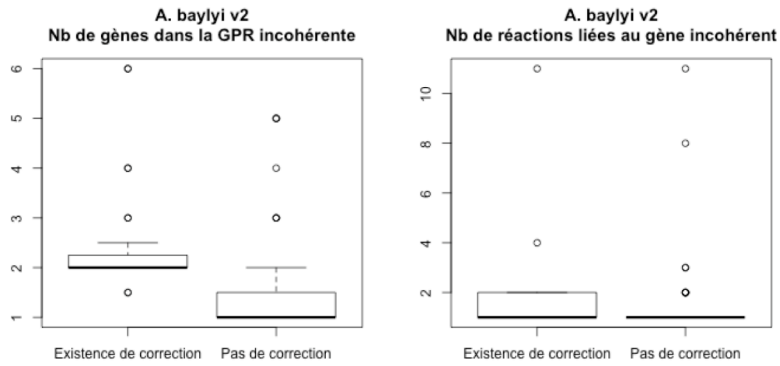
	Incohérence observée essentielle			Incohérence observée non essentielle			Tout type d'incohérence		
	cor.	non cor.	% cor.	cor.	non cor.	% cor.	cor.	non cor.	% cor.
A. baylyi v2									
succinate	25	29	*46%*	2	27	*7%*	27	56	*33%*
A. baylyi v3									
acetate	0	2	*0%*	2	7	*22%*	2	9	*18%*
asparagine	1	1	*50%*	1	7	*13%*	2	8	*20%*
butanediol	2	2	*50%*	3	8	*27%*	5	10	*33%*
glucarate	0	0	-	2	5	*29%*	2	5	*29%*
glucose	2	2	*50%*	1	9	*10%*	3	11	*21%*
lactate	0	2	*0%*	1	10	*9%*	1	12	*8%*
quinate	0	4	*0%*	2	8	*20%*	2	12	*14%*
urea	0	0	-	2	5	*29%*	2	5	*29%*
E. coli									
glucose	15	63	*19%*	7	22	*24%*	22	85	*21%*
glycerol	19	66	*22%*	6	23	*21%*	25	89	*22%*
B. subtilis									
rich	3	25	*11%*	11	7	*61%*	14	32	*30%*
S. cerevisiae									
ypd_ess	2	85	*2%*	0	0	-	2	85	*2%*
mmd	1	10	*9%*	3	13	*19%*	4	23	*15%*
ypd	3	70	*4%*	3	10	*23%*	6	80	*7%*
ypgal	2	0	*100%*	4	8	*33%*	6	8	*43%*
ypdge	3	61	*5%*	10	9	*53%*	13	70	*16%*
ypg	2	60	*3%*	5	9	*36%*	7	69	*9%*
ype	2	58	*3%*	5	16	*24%*	7	74	*9%*
ypl	1	60	*2%*	5	11	*31%*	6	71	*8%*

Tableau 19. Nombre d'incohérences pour lesquelles une correction GPR existe, pour chaque milieu pris séparément. Les résultats sont présentés en distinguant les incohérences des gènes observés essentiels, des incohérences des gènes observés non-essentiels. « cor. » une correction GPR existe ; « non cor. » aucune correction GPR n'existe.

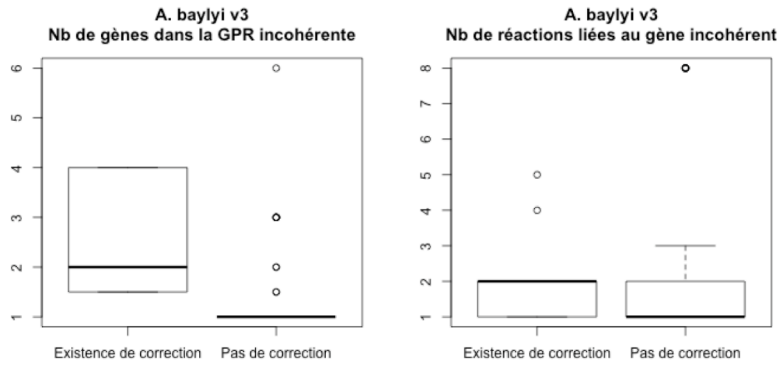
Dans l'ensemble, ce tableau montre que seule une minorité d'incohérences (entre 2% et 43%, selon le modèle et le milieu) pourrait être corrigée uniquement par des corrections GPR. Toutes les autres nécessitent de rechercher des corrections soit en dehors de la composante GPR soit en ajoutant de nouveaux gènes à associer aux réactions.

La répartition des corrections réalisables entre incohérence de gène essentiel et incohérence de gène non-essentiel est hétérogène entre les organismes. Alors que pour A. baylyi v2, une part élevée des incohérences de gènes essentiels dispose d'une correction GPR (46%), cette part est bien plus réduite pour les autres organismes (entre 2% et 22%, pour les milieux ayant plus de 10 incohérences). Inversement, peu d'incohérences de gènes non-essentiels disposent d'une correction chez A. baylyi v2 (7%), alors qu'entre 9% et 60% en disposent pour les autres organismes.

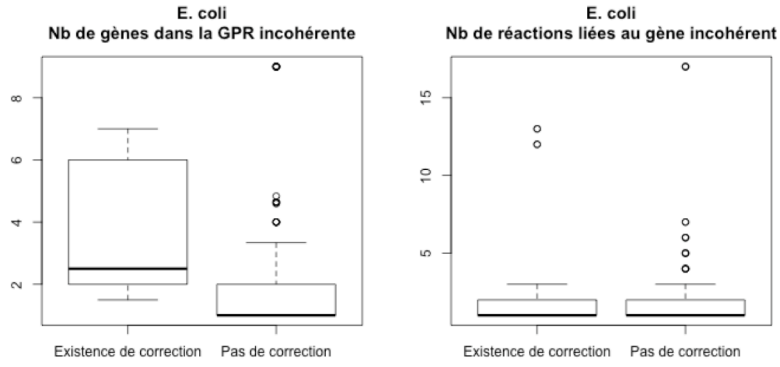
Afin d'évaluer l'effet de la taille des GPR et du nombre de réactions liées aux gènes sur l'existence d'une correction, nous avons tracé les Box Plots de ces deux grandeurs selon qu'une correction existe ou non (voir Figure 51).


A. baylyi v2
Nb de gènes dans la GPR incohérente
Existence de correction
Pas de correction

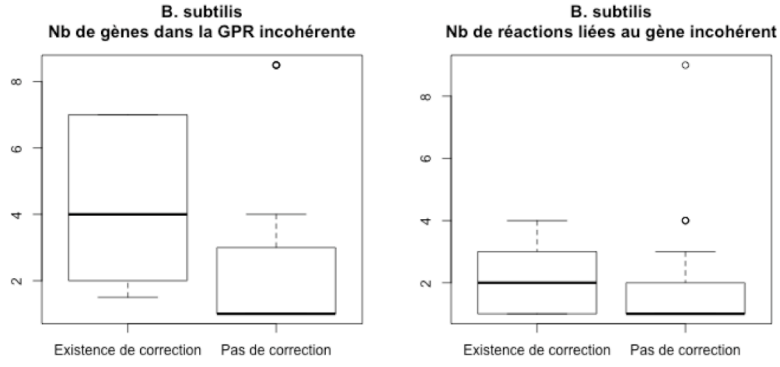
A. baylyi v2
Nb de réactions liées au gène incohérent
Existence de correction
Pas de correction


A. baylyi v3
Nb de gènes dans la GPR incohérente
Existence de correction
Pas de correction

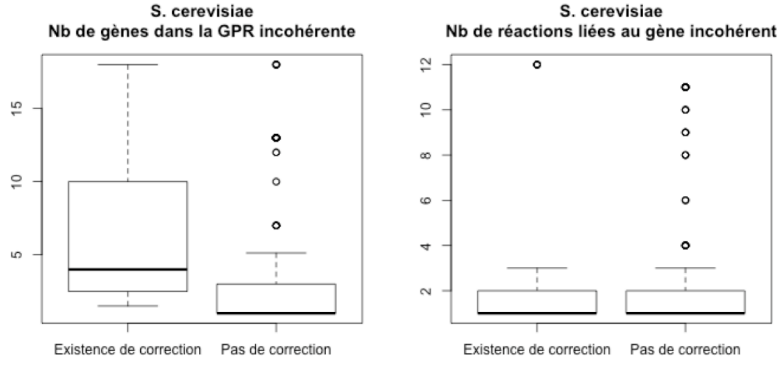
A. baylyi v3
Nb de réactions liées au gène incohérent
Existence de correction
Pas de correction


E. coli
Nb de gènes dans la GPR incohérente
Existence de correction
Pas de correction

E. coli
Nb de réactions liées au gène incohérent
Existence de correction
Pas de correction


B. subtilis
Nb de gènes dans la GPR incohérente
Existence de correction
Pas de correction

B. subtilis
Nb de réactions liées au gène incohérent
Existence de correction
Pas de correction


S. cerevisiae
Nb de gènes dans la GPR incohérente
Existence de correction
Pas de correction

S. cerevisiae
Nb de réactions liées au gène incohérent
Existence de correction
Pas de correction

Figure 51. Box Plot du nombre moyen de gènes contenus dans les GPR de chaque gène incohérent et du nombre de réactions liées à chaque gène incohérent, selon qu'une correction GPR existe ou non. Box plots de Tukey : chaque boîte est délimitée par les 1er et 3ème quartiles de la statistique et coupée par la médiane (ligne épaisse). Les moustaches représentent les valeurs minimales et maximales rencontrées, dans la limite de 1,5 fois l'écart interquartile. Les cercles représentent les valeurs sortant de ces moustaches (« outliers »).

Les incohérences pour lesquelles une correction GPR existe ont tendance à être reliées à des GPR initiales de tailles plus grandes que lorsque aucune correction n'existe (voir Figure 51 colonne de gauche). Cet effet s'interprète aisément, aucune correction ne pouvant naturellement être proposée par AutoGPR lorsqu'un seul gène est contenu dans la GPR à corriger. À l'inverse, on constate peu, voire aucune, différence entre les nombres de réactions reliées aux gènes incohérents disposant ou non de correction GPR. L'appartenance à une GPR de grande taille semble être ainsi le premier indicateur de l'existence d'une correction purement GPR.

Ce test d'existence de correction ne vérifie cependant pas la cohérence des corrections entre milieux. Pour cela, il est nécessaire d'exécuter la méthode complète AutoGPR.

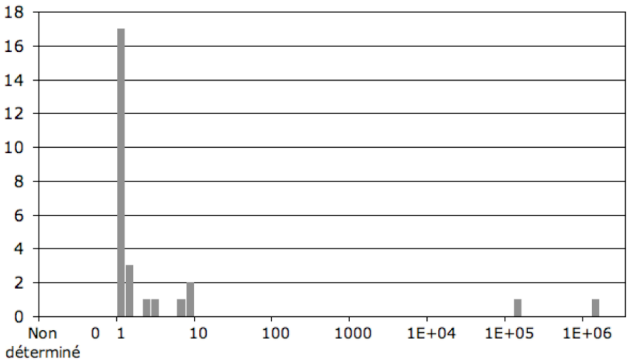
12.2.3 Proposition exhaustive de corrections GPR

Un préliminaire nécessaire à l'exécution d'AutoGPR est la détermination de tous les ensembles essentiels minimaux de réactions au sein des composantes connexes formées par les associations gènes-réactions (voir section 11.2.1). Nous les avons calculés pour tous les organismes et tous leurs milieux en suivant l'algorithme présenté précédemment. Dans la très grande majorité des cas, cet algorithme nous a permis de déterminer toutes les essentialités minimales de réactions. Celles-ci sont constituées pour la plupart de réactions individuelles, mais aussi de quelques rares ensembles de réactions « synthétiques létales ». Ces ensembles sont le plus souvent de taille 2, les quelques autres étant de taille 3 à 6. L'exécution de l'algorithme échoua sur les composantes connexes de 9 gènes d'E. coli et de 4 gènes de S. cerevisiae, tous milieux confondus ; ces composantes contiennent des ensembles essentiels de réactions trop grands pour être explorés par cet algorithme. Cependant, seul un de ces treize gènes est incohérent et ne pourra être traité par la suite.

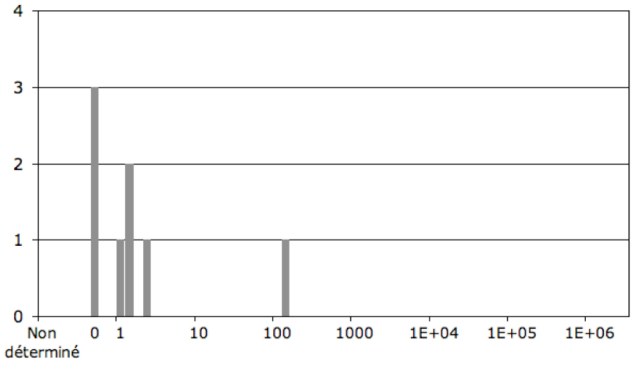
Enfin, dans un dernier temps, nous avons exécuté AutoGPR sur l'ensemble des incohérences disposant, selon le test d'existence, d'une correction GPR sur au moins

un des milieux. Des statistiques sur les corrections proposées sont données sur la Figure 52 ci-dessous.

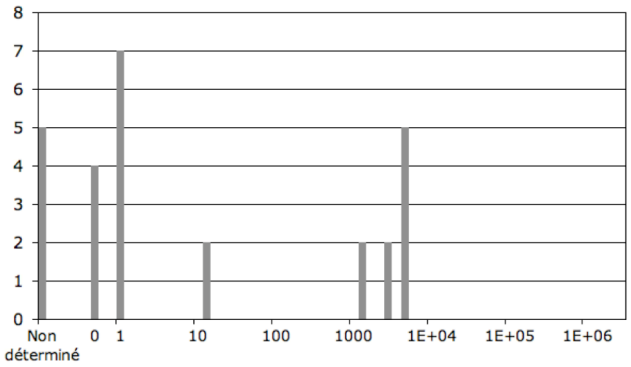
A. baylyi v2



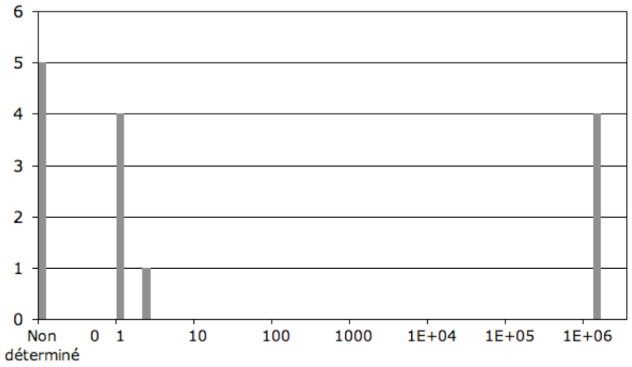
A. baylyi v3



E. coli



B. subtilis



S. cerevisiae

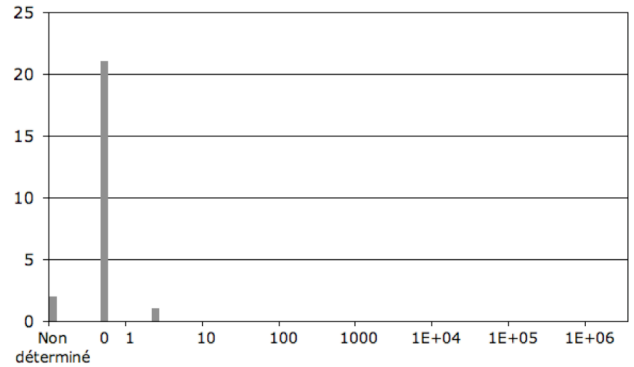


Figure 52. Histogrammes des nombres de GPR proposées par AutoGPR pour chaque incohérence identifiée précédemment comme ayant une correction sur chaque milieu pris séparément. Les GPR proposées sont compatibles simultanément sur tous les milieux. L'échelle des abscisses est en partie logarithmique : elle demeure linéaire entre chaque puissance de 10. Le cas « Non déterminé » indique les incohérences pour lesquelles AutoGPR n'a pu générer les corrections pour cause de nombre excessif de combinaisons.

Pour les modèles dont les données d'essentialités existent sur plusieurs milieux (A. baylyi v3, E. coli et S. cerevisiae), une part des incohérences ne dispose d'aucune correction GPR, alors même que des corrections étaient envisageables pour chaque milieu pris séparément. Ceci est particulièrement marqué chez S. cerevisiae où la quasi-totalité des incohérences ne dispose plus de correction. Ces cas révèlent des incompatibilités entre les contraintes d'essentialités posées sur les GPR par les différents milieux. Ces incompatibilités traduisent une différence de comportement de l'organisme entre les milieux qui n'est pas prise en compte par le modèle[74]. L'origine de l'incohérence peut dans ce cas se trouver dans les composantes RESEAU et BIOMASSE qui ne prédisent pas correctement les essentialités des réactions sur certains environnements. L'origine de l'incohérence peut également porter sur la régulation des gènes, certains d'entre eux ne s'exprimant que sur des milieux particuliers. Dans ce cas, l'hypothèse de « GPR constante sur tous les environnements » doit être levée pour pouvoir expliquer les phénotypes. Le recours à des règles de régulation dépendant de l'environnement qui modulent la présence des gènes dans les GPR – à l'image de la méthode rFBA (Covert et al. 2001) – semble être une solution permettant de prendre efficacement ces cas en compte (Covert et al. 2004). Nous évoquerons quelques cas précis de ce type dans la section suivante.

Lorsque des GPR compatibles existent, le nombre de combinaisons réalisables peut faire varier leur nombre sur plusieurs ordres de grandeur. Une forte proportion d'entre elles se limite néanmoins à une seule GPR compatible : dans ces cas, les données d'essentialités spécifient intégralement les GPR concernées (toujours bien entendu dans les limites des hypothèses d'AutoGPR). Dans les autres cas, le nombre de propositions demeure souvent limité (<10, notamment pour A. baylyi) mais peut rapidement atteindre des valeurs très élevées, jusqu'à rendre la recherche de GPR non réalisable (catégorie « Non déterminé » de la Figure 52). Cet effet est directement lié

[74] En supposant bien entendu que les expériences réalisées sur les différents milieux soient comparables et ne présentent pas de biais.

à la taille des GPR et au nombre de réactions liées aux gènes. Afin d'aider à la sélection des corrections GPR et rendre le calcul réalisable, les stratégies de sélection/simplification évoquées plus haut – distance à la GPR initiale et contrainte de GPR identiques pour plusieurs réactions – peuvent être employées (voir section 11.1).

Par exemple, l'incohérence du gène b0180 chez E. coli ne pouvait initialement être traitée par AutoGPR. Ce gène est lié à 12 réactions effectuant une activité 3-hydroxyacyl-ACP dehydratase pour 12 substrats distincts mais chimiquement proches du point de vue de cette activité.

Réaction	Equation bilan			GPR
3HAD40	3haACP[c]	->	but2eACP[c] + h2o[c]	(b0954 or b0180)
3HAD60	3hhexACP[c]	->	h2o[c] + thex2eACP[c]	(b0954 or b0180)
3HAD80	3hoctACP[c]	->	h2o[c] + toct2eACP[c]	(b0954 or b0180)
3HAD100	3hdecACP[c]	->	h2o[c] + tdec2eACP[c]	(b0954 or b0180)
3HAD120	3hddecACP[c]	->	h2o[c] + tddec2eACP[c]	(b0954 or b0180)
3HAD121	3hcddec5eACP[c]	->	h2o[c] + t3c5ddeceACP[c]	(b0180 or b0954)
3HAD140	3hmrsACP[c]	->	h2o[c] + tmrs2eACP[c]	(b0954 or b0180)
3HAD141	3hcmrs7eACP[c]	->	h2o[c] + t3c7mrseACP[c]	(b0954 or b0180)
3HAD160	3hpalmACP[c]	->	h2o[c] + tpalm2eACP[c]	(b0954 or b0180)
3HAD161	3hcpalm9eACP[c]	->	h2o[c] + t3c9palmeACP[c]	(b0180 or b0954)
3HAD180	3hoctaACP[c]	->	h2o[c] + toctd2eACP[c]	(b0954 or b0180)
3HAD181	3hcvac11eACP[c]	->	h2o[c] + t3c11vaceACP[c]	(b0180 or b0954)

Il est donc probable que ces réactions soient catalysées de manière similaire. En les contraignant à avoir toutes la même GPR, la déduction des GPR compatibles devient réalisable : seules deux configurations de GPR sont générées par AutoGPR[75].

Cette stratégie de simplification nécessite cependant d'entrer dans le détail des conversions chimiques catalysées par les gènes et ne peut ainsi être réalisée automatiquement. Nous proposerons d'autres pistes de simplification dans la section consacrée aux perspectives d'AutoGPR (section 13).

12.3 Comparaison des corrections d'AutoGPR aux interprétations expertes

Les incohérences de quatre des cinq modèles ont été examinées de manière experte par les auteurs de ces modèles. Il s'agit d'une part des modèles A. baylyi v2 et

[75] b0180 et b0954 sont tous deux essentiels, les réactions présentées ici sont essentielles. b0954 est en outre associé seul à une autre réaction essentielle. Les deux GPR proposées pour ces réactions sont donc « b0180 » et « b0180 ET b0954 ».

A. baylyi v3 pour lesquels nous avions explicitement déterminé des corrections (voir article inclus section 8) et d'autre part des modèles B. subtilis et E. coli dont les auteurs ont proposés des interprétations à chaque incohérence. Cette section évalue la pertinence des propositions d'AutoGPR en les confrontant à ces corrections et interprétations.

12.3.1 Comparaison aux corrections des modèles d'*A. baylyi*

Lors du processsus de raffinement du modèle d'*A. baylyi*, nous avions interprété manuellement les incohérences et proposé des corrections dans les composantes GPR, RESEAU et BIOMASSE du modèle. Le Tableau 20 ci-dessous récapitule l'ensemble des incohérences détectées pour les modèles v2 et v3 ainsi que les corrections et interprétations réalisées pour chacune d'entre elles, et les compare aux propositions de correction d'AutoGPR.

CORRECTION

	gène		ess	test	autogpr n	cor	sel
GPR							
activité nécessitant simultanément tous les gènes							
2	ACIAD0661	hisG	E	o	1	o	P
2	ACIAD1257	hisZ	E	o	1	o	P
2	ACIAD3103	ilvH	E	o	8	o	SP
gènes associés à une autre réaction							
2	ACIAD2606		E	-	-		
isozyme non fonctionnelle							
2	ACIAD0151	guaA	E	o	1	o	P
2	ACIAD0249	ribC	E	o	2	o	P
2	ACIAD0871	fabG	E	o	2E+05	o	SP
2	ACIAD1069	lysS	E	o	1	o	P
2	ACIAD1255	epd	E	o	9	o	
2	ACIAD1323	purF	E	-	-		
2	ACIAD1375	cdsA	E	o	1	o	P
2	ACIAD1736	accC	E	o	1	o	P
2	ACIAD1737	accB	E	o	1	o	P
2	ACIAD1925	fda	E	o	1	o	P
2	ACIAD2227	dctA	E	o	1E+06	o	P
2	ACIAD2565	gap	E	o	9	o	
2	ACIAD2666		E	o	1	o	P
2	ACIAD2907	prs	E	o	1	o	P
2	ACIAD3062	folK	E	o	1	o	P
2	ACIAD3249	ribA	E	o	1	o	P
2	ACIAD3365	murE	E	o	1	o	P
2	ACIAD3371	gltX	E	o	4	o	P
3	ACIAD1710	pcaC	E	-	-		
3	ACIAD2018	ald1	E	o	0	n	
3	ACIAD2088	aspQ	E	o	4	o	P
3	ACIAD2983	gcd	E	o	2	o	P
présence d'un enzyme alternative							
2	ACIAD1231	argD	D	-	-		
2	ACIAD1642	uppP	D	-	-		
2	ACIAD2968	ispA	D	-	-		
3	ACIAD1020	acoD	D	o	1	n	
3	ACIAD1715	quiX	D	-	-		
3	ACIAD2984		D	-	-		
réaction occurant spontanément							
3	ACIAD2819		D	-	-		
fausse sous-unité d'un complexe							
2	ACIAD0799		D	o	1	n	
RESEAU							
fausse voie alternative							
2	ACIAD0239	ppa	E	-	-		
2	ACIAD0547	proA	E	-	-		
2	ACIAD1105	adk	E	-	-		
2	ACIAD1920	glnS	E	-	-		
2	ACIAD2560	proB	E	-	-		
2	ACIAD3032	proC	E	-	-		
voie alternative manquante							
2	ACIAD0106	lldP	D	-	-		
2	ACIAD0451	katA	D	-	-		
2	ACIAD0901	dut	E	-	-		
2	ACIAD0930	glpK	D	-	-		
2	ACIAD1045	metH	D	-	-		
3	ACIAD0106	lldP	D	-	-		
BIOMASSE							
précurseur de biomasse non essentiel							
2	ACIAD0076	rmlB	D	-	-		
2	ACIAD0078	rmlD	D	-	-		
2	ACIAD0079	rmlA	D	-	-		
2	ACIAD0080	rmlC	D	-	-		
2	ACIAD0086	epsM	D	-	-		
2	ACIAD0099	galU	D	-	-		
2	ACIAD0101	pgi	D	-	-		
2	ACIAD0104	manB	D	-	-		
2	ACIAD2429	cyoE	D	-	-		
précurseur de biomasse manquant							
2	ACIAD1374	ispU	E	-	-		

PAS DE CORRECTION

	gène		ess	test	autogpr n	cor	sel
Interprétation validée							
erreur expérimentale							
3	ACIAD0108	lldD	D	-	-		
manque connu dans la connaissance d'une voie							
2	ACIAD0856	bioA	E	-	-		
2	ACIAD0857	bioF	E	-	-		
2	ACIAD0859	bioD	E	-	-		
2	ACIAD2045	bioB	E	-	-		
auxotrophie non modélisée							
2	ACIAD3523	metE	E	-	-		
Interprétation hypothétique							
2	ACIAD0556	ndk	D	-	-		
2	ACIAD0650	argJ	E	o	2	n	
2	ACIAD1150	pyrC	E	o	1	n	
2	ACIAD1346	sodB	E	-	-		
2	ACIAD2282	sahH	D	-	-		
2	ACIAD2314	metZ	E	o	2	n	
2	ACIAD2458	glnA	E	o	1	n	
2	ACIAD2842	pckG	E	-	-		
2	ACIAD2847	folD	E	o	5	n	
2	ACIAD3155	mdh	E	-	-		
2	ACIAD3349	gltD	E	-	-		
2	ACIAD3350	gltB	E	-	-		
2	ACIAD3470	msuE	E	-	-		
2	ACIAD3506	aceF	E	-	-		
3	ACIAD0546		E	-	-		
3	ACIAD0556	ndk	D	-	-		
3	ACIAD1021		D	-	-		
3	ACIAD1707	pcaB	E	-	-		
3	ACIAD1711	pcaH	E	-	-		
3	ACIAD1712	pcaG	E	-	-		
3	ACIAD1744	aspA	E	-	-		
Pas d'interprétation précise							
2	ACIAD0072	ugd	E	-	-		
2	ACIAD0173	rhtB	E	-	-		
2	ACIAD0382	ubiB	D	-	-		
2	ACIAD0505	purU1	E	-	-		
2	ACIAD1482	kdsD	D	-	-		
2	ACIAD1483	kdsC	D	-	-		
2	ACIAD2283	metF	D	-	-		
2	ACIAD2290	cydA	E	-	-		
2	ACIAD2525		E	-	-		
2	ACIAD2667	pdxB	D	-	-		
2	ACIAD2788		E	-	-		
2	ACIAD2880	sdhA	D	o	1	n	
2	ACIAD2911	panD	D	-	-		
2	ACIAD3503	guaB	E	-	-		
2	ACIAD3510	lpxC	D	-	-		
3	ACIAD0086	epsM	E	-	-		
3	ACIAD0382	ubiB	D	-	-		
3	ACIAD0922		E	-	-		
3	ACIAD2070	metI	E	-	-		
3	ACIAD2282	sahH	D	-	-		
3	ACIAD2283	metF	D	o	2	n	
3	ACIAD2667	pdxB	D	-	-		
3	ACIAD2755		E	o	0	n	
3	ACIAD2875	sucB	E	-	-		
3	ACIAD2876	sucA	E	-	-		
3	ACIAD2880	sdhA	D	o	147	n	
3	ACIAD2911	panD	E	-	-		
3	ACIAD3071	cysM	E	o	0	n	
Incohérence ajoutée par la méthode FBA							
2	ACIAD2456	ubiC	D	-	-		
2	ACIAD3383	acr1	D	-	-		
3	ACIAD0080	rmlC	D	-	-		
3	ACIAD0099	galU	D	-	-		
3	ACIAD0104	manB	D	-	-		
3	ACIAD3383	acr1	D	-	-		
3	ACIAD3549	gshA	E	-	-		

Tableau 20. Comparaison des propositions d'AutoGPR aux corrections et interprétations des incohérences des modèles A. baylyi v2 et A. baylyi v3. Les incohérences sont classées par type de correction et d'interprétation. La colonne de gauche identifie le modèle concerné par l'incohérence (2 ou 3). Signification des colonnes : ess, essentialité de l'incohérence (sur le profil de milieux pour A. baylyi v3 ; E essentiel, D non-essentiel) ; test, résultat du test d'existence de correction GPR (o existence de correction) ; n, nombre de propositions d'AutoGPR ; cor, présence de la correction experte dans les propositions d'AutoGPR (o oui, n non) ; sel, méthode de sélection de la GPR (P la plus proche de la GPR initiale, S réactions contraintes à avoir des GPR similaires).

Sensibilité[76]

Parmi les 34 incohérences que nous avons corrigées dans la composante GPR (pour les deux modèles), 24 disposent de propositions d'AutoGPR. Pour 22 d'entre elles, la correction appliquée est incluse dans les propositions, donnant un score global de sensibilité de 65% pour les modèles A. baylyi.

AutoGPR propose dans la majorité des cas des corrections uniques aux incohérences. Lorsque plusieurs corrections distinctes sont suggérées, les stratégies de sélection permettent d'identifier efficacement la correction retenue. En effet, en contraignant dans deux cas des réactions à avoir des GPR similaires (ACIAD3103 et ACIAD0871, cas analogues à celui présenté pour E. coli section 12.2.3), la correction retenue correspond dans 7 cas sur 9 à la proposition d'AutoGPR la plus proche des GPR initiales. Ces stratégies peuvent ainsi s'avérer être des outils utiles à la sélection des GPR les plus probables.

Seuls deux types de corrections GPR (toutefois majoritaires) sont pris en compte par AutoGPR : la détection (1) d'isozyme non fonctionnelle et (2) d'activités nécessitant la présence de tous les gènes (sous-unités de complexes enzymatiques).

Pour le premier type de correction, AutoGPR propose dans 86% des cas (19/22) la correction retenue. Il s'agit le plus souvent de retirer une isozyme hypothétique d'une GPR afin de retrouver l'essentialité de l'enzyme principale. Dans le cas par exemple des incohérences de *epd* (ACIAD1255) et *gap* (ACIAD2565) qu'AutoGPR corrigea

[76] Nous utiliserons un peu par abus de langage les termes de sensibilité et de spécificité pour désigner respectivement la part de corrections expertes GPR retrouvées par AutoGPR et la part de corrections d'AutoGPR effectivement retenues. AutoGPR déduisant toutes les corrections GPR réalisables dans son champ d'application, sa spécificité est théoriquement de 100%. La « spécificité » que nous utiliserons ici cherche plutôt à évaluer quelle part des corrections expertes de la composante GPR rentre dans le champ d'application d'AutoGPR.

correctement, la correction consiste à associer chaque gène spécifiquement à une réaction alors qu'ils étaient considérés initialement comme des isozymes de ces réactions. Les trois cas où AutoGPR ne proposa pas de solution sont dus soit à la présence d'une autre incohérence non résoluble dans la même composante connexe (ACIAD1323) soit à la nécessité d'effectuer en plus une correction dans la composante RESEAU du modèle (ACIAD1710 et ACIAD2018).

Pour le deuxième type de correction, AutoGPR proposa dans les trois cas la correction experte réalisée. Parmi les corrections de ce type, celle de l'incohérence de *ilvH* (ACIAD3103) mérite d'être détaillée (voir Figure 53).

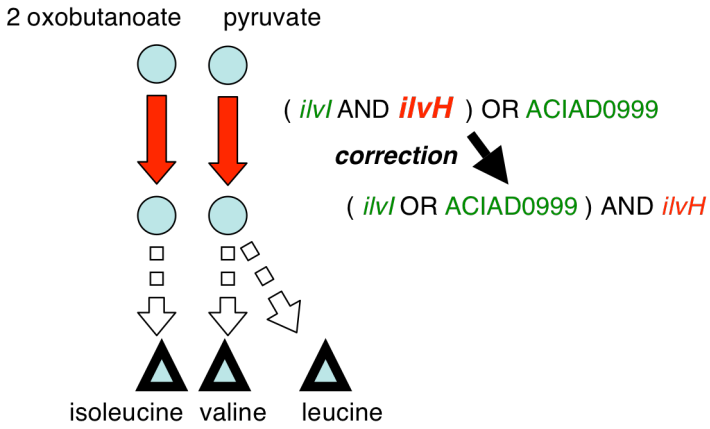


Figure 53. Correction de l'incohérence du gène *ilvH* dans A. baylyi v2. La couleur rouge (respectivement verte) indique une réaction ou un gène essentiel (respectivement non-essentiel). Le gène en gras possède une essentialité incohérente avec la prédiction du modèle. Les métabolites en triangle sont des précurseurs de biomasse.

Initialement, deux isozymes étaient supposés catalyser deux réactions essentielles d'activité de type acetolactate synthase : un complexe IlvI & IlvH et une protéine codée par ACIAD0999. Cette redondance génétique contredisant l'essentialité de *ilvH*, AutoGPR génère 8 corrections distinctes qui combinent indépendamment des impacts de la délétion de *ilvH* sur chaque réaction. En contraignant les deux réactions à avoir des GPR identiques, seules deux corrections demeurent : « *ilvH* » ou « (*ilvI* OU ACIAD0999) ET *ilvH* ». L'examen de la littérature nous avait permis de déterminer cette deuxième correction : *ilvH* est en réalité un facteur de stabilisation pouvant fonctionner indistinctement avec les unités catalytiques alternatives codées par ACIAD0999 ou *ilvI* (Vyazmensky et al. 1996). Cet exemple illustre le fait que les données d'essentialités (sur lesquelles AutoGPR repose) peuvent dans certains cas aider à déterminer de manière précise des règles d'associations complexes entre gènes et réactions.

Les quatre autres types de corrections GPR expertes appliquées aux modèles d'*A. baylyi* impliquent d'ajouter de nouveaux gènes aux GPR. Ces corrections sortent de ce fait du cadre d'application d'AutoGPR, expliquant la quasi-inexistence de proposition pour les incohérences de ces types.

Il est intéressant de noter ici que toutes les corrections GPR correctement détectées par AutoGPR concernent des gènes essentiels. Il semble donc sur cet exemple qu'AutoGPR ait un biais très fort envers les corrections d'incohérences de gène essentiel. Ce biais peut s'interpréter de la manière suivante. La correction type d'AutoGPR pour les gènes non-essentiels incohérents (prédits donc essentiels) consiste à remplacer une relation ET en OU, c'est à dire à considérer les sous-unités d'un complexe comme isozymes. Les complexes faisant généralement l'objet d'une attention particulière lors de leur construction, il est relativement rare d'introduire par erreur des interactions entre sous-unités. Toute autre correction GPR d'une incohérence de gène non-essentiel consiste à ajouter une isozyme en introduisant un nouveau gène dans la GPR (voir Tableau 20). Ce type de modification sort alors du cadre d'action d'AutoGPR (correction à ensemble de gènes constant).

Spécificité

S'agissant de la spécificité de la méthode, sur 32 incohérences pour lesquelles AutoGPR propose des corrections, 22 ont été corrigées manuellement selon une des propositions d'AutoGPR (score de spécificité de 69%). Les corrections AutoGPR de deux incohérences sont explicitement en désaccord avec la correction réelle (ACIAD1020 et ACIAD0799), les corrections appliquées nécessitant d'ajouter un nouveau gène à la GPR.

Il est cependant important de considérer ce score avec précaution. En effet, certaines corrections d'AutoGPR non retenues peuvent toutefois avoir un réel sens biologique et aider à interpréter les incohérences. Ainsi, les incohérences des gènes essentiels *pyrC* et *folD* sont résolues par AutoGPR en supprimant leurs isozymes. Cependant, aucune information supplémentaire ne permettant de corroborer ni d'infirmer ces corrections, celles-ci n'ont pas été réalisées et nécessitent d'investiguer plus avant les activités des isozymes. Dans un autre ordre d'idée, l'incohérence du gène essentiel *glnA* (ACIAD2458), qui catalyse l'activité glutamine synthetase conjointement avec le gène *glnT* (ACIAD2528), pourrait s'expliquer par l'existence

d'une contrainte régulatrice favorisant l'utilisation de l'un ou l'autre de ces gènes en fonction de la disponibilité en ammonium (Reitzer 1996). Pour cette incohérence, AutoGPR propose justement d'écarter *glnT* de la GPR de la glutamine synthetase, traduisant le fait que l'enzyme de ce gène n'est potentiellement pas active dans les conditions des expériences.

Il est ainsi probable que, dans l'ensemble des incohérences n'ayant pu être interprétées, les propositions réalisées par AutoGPR puissent fournir des pistes d'investigation utiles.

12.3.2 Comparaison aux interprétations expertes des modèles de *B. subtilis* et *S. cerevisiae*

Bien que les incohérences des modèles B. subtilis et S. cerevisiae n'aient pas donné lieu à des corrections, leurs auteurs les ont tout de même examinées de manière experte afin d'en déterminer la cause. L'ensemble de ces interprétations étant mis librement à disposition sous la forme de données supplémentaires aux articles des modèles (Oh et al. 2007; Duarte, Herrgard et al. 2004), nous les avons utilisées pour évaluer la pertinence des propositions d'AutoGPR pour ces modèles.

B. subtilis

Le Tableau 21 confronte les interprétations expertes réalisées pour B. subtilis aux propositions d'AutoGPR pour ce modèle.

gène	ess	autogpr test	n	Commentaire des auteurs
GPR				
BG10412 fbaA	E	-	-	Possibly regulation effect. fbaA could not be replaced by fbaB.
BG11955 murAA	E	o	1	Possibly regulation effect. murAA could not replaced by a homologue murAB.
BG11964 racE	E	o	1	Regulation effect. racE could not be replaced by yrpC.
BG12391 yumC	E	-	-	Possibly regulation effect. Both yumC and trxB products seem to be essential for cell growth.
BG12398 trxB	E	-	-	Possibly regulation effect. Both yumC and trxB products seem to be essential for cell growth.
RESEAU				
BG10282 ndk	D	-	-	Metabolic gap and/or external source
BG10305 bkdB	D	o	1E+06	Metabolic gap and/or external source
BG10306 bkdAB	D	o	1E+06	Metabolic gap and/or external source
BG10307 bkdAA	D	o	1E+06	Metabolic gap and/or external source
BG11725 lpdV	D	o	1E+06	Metabolic gap and/or external source
BG13951 yubB	D	-	-	Metabolic gap and/or external source
BIOMASSE				
BG10402 gtaB	D	-	-	Biomass composition
BG10548 dltD	D	o	ND	Biomass composition
BG10549 dltC	D	o	ND	Biomass composition
BG10550 dltB	D	o	ND	Biomass composition
BG10551 dltA	D	o	ND	Biomass composition
BG10724 tagE	D	o	ND	Biomass composition
BG11012 pssA	D	-	-	Biomass composition
BG11013 psd	D	-	-	Biomass composition
BG11192 ggaB	D	o	1	Biomass composition
BG11367 ggaA	D	o	1	Biomass composition
BG11611 ugtP	D	-	-	Biomass composition
BG11840 metK	E	-	-	Synthesize S-adenosylmethionine, which is necessary for siroheme synthesis.
BG12089 acpS	E	-	-	Synthesize acyl-carrier protein.
BG12900 yfiX	D	-	-	Biomass composition
BG13824 ytaG	E	-	-	Synthesize coenzyme A.
Autre interprétation				
BG10897 tpiA	E	-	-	Possibly toxic effect. Accumulation of dihydroxyacetone phosphate, which may leads to the formation of the bactericidal compound methylglyoxal
BG11062 pgk	E	-	-	Conditionally essential gene.
BG11937 fmt	E	-	-	Other function. Required for the formylation of methionyl tRNA
BG12344 mrpF	E	-	-	Other function. Multiple resistance and pH homeostasis
BG12345 mrpD	E	-	-	Other function. Multiple resistance and pH homeostasis
BG12355 mrpA	E	-	-	Other function. Multiple resistance and pH homeostasis
BG12356 mrpB	E	-	-	Other function. Multiple resistance and pH homeostasis
BG12357 mrpC	E	-	-	Other function. Multiple resistance and pH homeostasis
BG13966 yueK	E	-	-	Toxic effect. Nicotinate accumulation.
Pas d'interprétation précise				
BG10073 guaB	E	-	-	Not well understood.
BG10131 hprT	E	-	-	Not well understood.
BG10207 pdhA	E	-	-	Not well understood.
BG10273 odhB	E	-	-	Not well understood.
BG10410 pyrG	E	-	-	Not well understood.
BG11004 cmk	E	o	4	Not well understood.
BG11247 tkt	E	-	-	Not well understood.
BG11404 nrdE	E	-	-	Not well understood.
BG11405 nrdF	E	-	-	Not well understood.
BG11426 ymaA	E	-	-	Not well understood.
BG12644 pfkA	E	-	-	Not well understood.

Tableau 21. Comparaison des propositions d'AutoGPR aux interprétations expertes des incohérences de B. subtilis. Les gènes incohérents sont classés par type d'interprétation (les commentaires des auteurs sont repris en dernière colonne). Signification des colonnes : ess, essentialité du gène (E essentiel, D non-essentiel) ; test, résultat du test d'existence de correction GPR (o existence de correction) ; n, nombre de propositions d'AutoGPR (ND proposition non réalisable).

Nous avons pu classer les commentaires des auteurs en cinq catégories, selon que l'interprétation se rattache à une des composantes du modèle (GPR, RESEAU et BIOMASSE) ou non (« Autre interprétation » et « Pas d'interprétation précise »).

Toutes les interprétations liées à la composante GPR supposent ici l'existence d'interactions régulatrices inhibant l'expression d'isozymes. Sur les cinq incohérences de ce type, deux disposent d'une proposition d'AutoGPR correspondant correctement à l'interprétation. Pour les trois cas ne correspondant pas (*fbaA*, *yumC* et *trxB*), les réactions catalysées ne sont en fait elles-mêmes pas essentielles : une correction purement GPR ne peut donc pas exister (aucune réaction essentielle à impacter par ces gènes essentiels). Pour corriger ces incohérences selon les interprétations avancées ici, il est nécessaire d'effectuer également des corrections aux composantes RESEAU ou BIOMASSE afin de rendre ces réactions essentielles. Il est intéressant de remarquer pour cet organisme aussi que l'ensemble des interprétations de type GPR concernent des incohérences de gènes essentiels, corroborant la remarque faite pour A. baylyi.

AutoGPR propose des corrections GPR pour une part significative des incohérences interprétées comme liées aux composantes RESEAU et BIOMASSE. Tous ces cas correspondent à des incohérences de gènes non-essentiels appartenant à des complexes. Pour rendre ces gènes effectivement non-essentiels, AutoGPR propose logiquement de transformer les relations de « complexes » en relations d'« isozymes ». Cependant, l'examen des fonctions de ces gènes montre clairement que ceux-ci ne peuvent se comporter comme des isozymes et forment réellement un ensemble nécessitant leur présence conjointe. Si AutoGPR était contraint à ne pas effectuer ce type de transformation pour ces complexes, aucune correction GPR n'aurait alors été proposée. Ceci corrobore également une autre remarque faite ci-dessus pour A. baylyi à propos de la correction de gènes non-essentiels.

En conclusion pour B. subtilis, la correspondance entre les interprétations et les propositions d'AutoGPR est donc fortement altérée par ce comportement d'AutoGPR sur les complexes. En revanche, les deux seules interprétations purement GPR (*racE* et *yumC*) sont correctement détectées par AutoGPR.

S. cerevisiae

Avec des interprétations pour plus de 230 incohérences, le modèle S. cerevisiae offre une perspective de confrontation avec les propositions d'AutoGPR plus large que pour les modèles précédents. Cependant, les incompatibilités entre milieux font qu'une seule incohérence dispose d'une correction GPR compatible simultanément avec les phénotypes des huit milieux de S. cerevisiae. La présence de régulation, les conditions expérimentales différentes entre les études et la possibilité d'erreurs dans les composantes RESEAU et BIOMASSE peuvent expliquer ces incompatibilités.

Les auteurs n'ont cependant pas cherché à interpréter systématiquement les incohérences sur tous les milieux à la fois. Nous avons donc plutôt confronté leurs interprétations aux résultats des tests d'existence de correction GPR, valables pour chaque milieu pris séparément. Le Tableau 22 présente les résultats de cette confrontation regroupés par catégorie d'interprétation, tels que définis par les auteurs eux-mêmes.

Catégorie d'interprétation		**cor.**	**non cor.**	**% cor.**
Acc	Accumulation d'intermédiaire toxique	0	6	0%
Bio	Problème avec la constitution de la biomasse	1	41	2%
Den	Impasse métabolique dans le modèle	0	8	0%
Dis	Incohérence entre données expérimentales	2	12	14%
Iso	Problème avec les associations GPR	10	9	**53%**
Med	Problème dans la composition du milieu *in silico*	7	22	24%
Mod	Problème dans la structure du réseau métabolique	0	4	0%
Oth	Implication du gène dans des processus non métaboliques	1	78	1%
Reg	Régulation transcriptionnelle manquante	2	0	**100%**
Slo	Croissance ralentie *in silico*	0	4	0%
Unk	Cause de fausse prédiction inconnue	1	28	3%
	Incohérence non présente dans l'article	*0*	*1*	*0%*

Tableau 22. Comparaison des types d'interprétation des incohérences de S. cerevisiae aux tests d'existence de correction GPR. Rappel : ces tests ne vérifient l'existence de correction GPR que sur chaque milieu pris séparément, les corrections n'étant pas nécessairement compatibles entre les milieux. Signification des colonnes : cor, nombre d'incohérences disposant (selon le test) d'une correction GPR sur chaque milieu ; non cor, nombre d'incohérences ne disposant pas de correction GPR sur au moins un des milieux ; % cor, part des incohérences disposant d'une correction GPR sur chaque milieu (cor/(cor + non cor)). Les % en gras indiquent les cas où les incohérences disposant d'une correction GPR sont majoritaires.

Les incohérences disposant d'après AutoGPR de corrections GPR sont majoritaires dans les catégories Reg et Iso, et minoritaires (souvent largement) dans toutes les autres. Les catégories Reg et Iso sont les seules à concerner explicitement la

composante GPR, les tests d'existences de correction GPR correspondent ainsi globalement bien aux classes interprétations.

La catégorie Reg regroupe deux incohérences dues à des régulations. Toutes deux ont été clairement interprétées comme dépendant de l'environnement : la première implique une pyruvate kinase (*CDC19* [77]) dont la seule isozyme (*PYK2* [78]) est connue pour être active uniquement pour de faibles flux glycolytiques, rendant *CDC19* essentielle sur milieu glucose ; la seconde implique la réaction acetaldehyde dehydrogenase dont seule une des cinq isozymes (*ADH1* [79]) est a priori exprimée sur milieu glucose, celui-ci réprimant l'expression des autres. AutoGPR identifie correctement l'existence de correction sur les milieux glucose (ypd et ypd_ess). Cependant, ces corrections sont incompatibles avec les autres milieux, car les isozymes y « redeviennent » actives.

La catégorie Iso regroupe des interprétations liées aux GPR elles-mêmes, incluant majoritairement l'ajout ou le retrait d'isozyme et la détection de sous-unité non-essentielle dans un complexe. Sur les 19 incohérences classées dans cette catégorie, 10 disposent de corrections individuellement sur les milieux selon AutoGPR. Parmi ces dernières 8 sont des incohérences de gènes essentiels et 2 de gènes non-essentiels, corroborant ici aussi le biais constaté pour A. baylyi et S. cerevisiae. Les 9 incohérences ne disposant pas de corrections sont quant à elles équiréparties entre gènes essentiels et non-essentiels.

Tous les autres types d'interprétations ne concernent pas la composante GPR ; il est donc naturel que l'existence de correction selon AutoGPR y soit minoritaire. Dans ces catégories, 12 incohérences disposent tout de même d'une correction GPR. 11 d'entre elles concernent des gènes non-essentiels. Ce résultat corrobore ici aussi un constat effectué avec B. subtilis, selon lequel les propositions de correction pour les gènes non-essentiels ont tendance à être moins réalistes que celles pour les gènes essentiels.

[77] Nom systématique : YAL038W
[78] Nom systématique : YOR347C
[79] Nom systématique : YOL086C

13 Limites et perspectives

En conclusion, le développement de la méthode AutoGPR nous a montré que, exploitées de manière appropriée, les données d'essentialité pouvaient conduire par des déductions logiques à proposer automatiquement des corrections aux relations GPR. L'implémentation que nous avons retenue ici – déduction systématique de toutes les GPR envisageables, à ensemble constant de gènes – nous a permis d'illustrer l'intérêt et de montrer la faisabilité d'une telle méthode sur cinq modèles. Nous avons cependant évoqué à plusieurs reprises l'existence de limites et de faiblesses. Nous allons les reprendre ici et proposer des possibilités d'amélioration de la méthode, ainsi que des perspectives d'utilisation plus large.

13.1 Réduction de la combinatoire des propositions de correction

Lors de la génération des corrections GPR pour les cinq modèles, nous avions constaté que le nombre de propositions pouvaient devenir particulièrement élevé, en particulier lorsque les composantes connexes comptaient plusieurs gènes et réactions. Cet effet, dû aux combinaisons des différents impacts possibles pour chaque délétion de gène, s'avère particulièrement gênant. D'une part, il augmente le nombre de propositions à considérer et, d'autre part, rend dans certains cas les déductions tout bonnement impossibles.

Cependant, un examen plus approfondi des spécifications sur les GPR déduites des scénarios d'impacts met en évidence des spécifications non-informatives qui augmentent inutilement le nombre d'alternatives. La Figure 54 illustre cet effet sur l'exemple que nous avions utilisé dans la partie théorique.

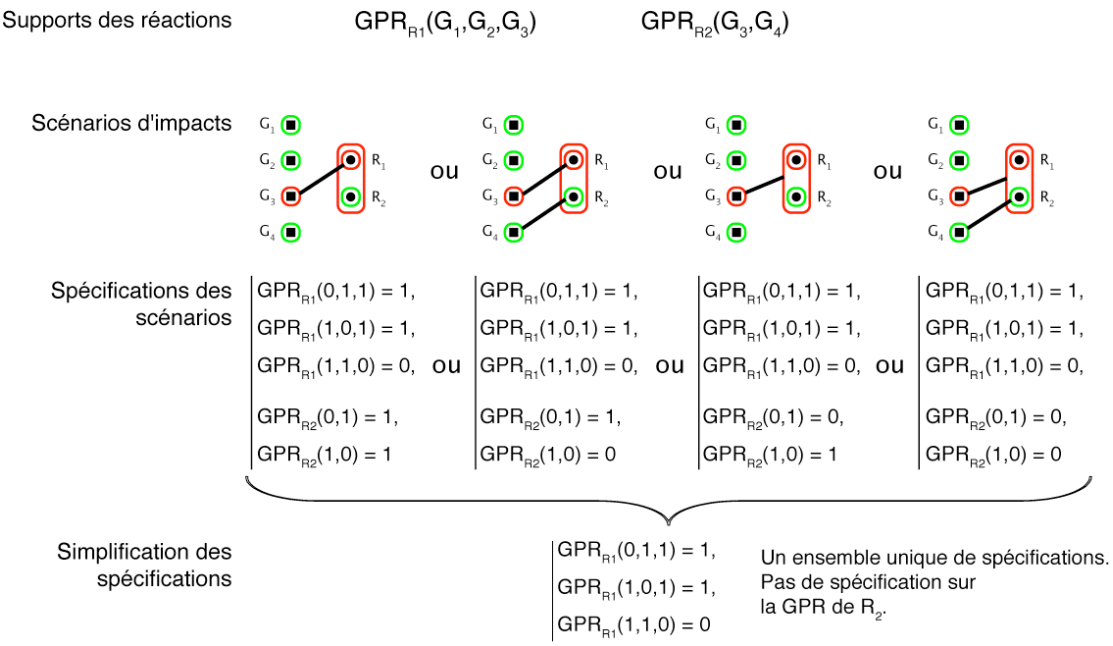


Figure 54. Simplification des spécifications issues des scénarios d'impact générés par AutoGPR. Les différents scénarios proposent des spécifications alternatives pour R_2 qui, lorsqu'on considère tous les ensembles de spécifications simultanément, se simplifient. Seul un ensemble de spécifications se déduit réellement des contraintes posées par les essentialités et les supports de réactions.

Les quatre scénarios d'impact générés par AutoGPR se traduisent en quatre ensembles alternatifs de spécifications sur R_1 et R_2. En examinant simultanément ces quatre ensembles, il s'avère que les spécifications alternatives pour R_2 décrivent en réalité la totalité de ses comportements possibles. De ce fait, les quatre ensembles se simplifient pour ne retenir que les spécifications sur R_1, seules spécifications réellement à l'œuvre. Lors de l'étape d'implémentation des GPR, ces spécifications détermineront les GPR correctes pour R_1, la GPR initiale de R_2 pouvant être conservée pour R_2.

En incorporant directement ce type de simplification dans la partie spécification de la méthode AutoGPR, le nombre de propositions « non-informatives » pourrait alors se trouver largement réduit dans certains cas.

13.2 Amélioration de la spécificité pour les corrections de gènes non-essentiels

Alors qu'AutoGPR détecte avec une relative bonne spécificité des corrections GPR pour les gènes essentiels, nous avons relevé que la spécificité pour les gènes non-essentiels est bien moins bonne. Comme déjà évoqué plus haut à propos du modèle B. subtilis (voir 12.3.2), ceci est principalement dû au fait que les corrections

proposées par AutoGPR pour les gènes non-essentiels consistent à transformer des sous-unités d'un complexe en isozymes, corrections relativement peu probables étant donné l'attention particulière généralement portée à la construction des complexes.

La spécificité d'AutoGPR pourrait être améliorée en prenant en compte cette information sur la fiabilité des complexes. Pour chaque complexe dont les interactions entre sous-unités sont confirmées (ou lorsque la probabilité d'existence de l'interaction dépasse un seuil de confiance), une méthode interdisant à AutoGPR de remplacer la relation ET en relation OU entre les gènes permettrait d'éliminer ces cas de fausse proposition.

13.3 Au delà des trois hypothèses fondamentales d'AutoGPR

Les trois hypothèses fondamentales sur lesquelles repose AutoGPR – associations gènes-réactions connues, composantes RESEAU et BIOMASSE fixes, GPR identiques sur tous les milieux – définissent précisément son champ d'action, mais, en contrepartie, écartent d'autres types d'interprétations. Nous discuterons rapidement de ces cas dans cette partie en proposant des pistes d'amélioration.

13.3.1 Associations gène-réaction prédéfinies

L'hypothèse d'associations gènes-réactions prédéfinies prive AutoGPR d'une part significative des corrections réalisées dans la composante GPR, dans lesquelles de nouveaux gènes sont ajoutés aux GPR.

La recherche de gènes candidats pouvant être nouvellement associés à des réactions est un thème de recherche à part entière, qui sort du cadre d'AutoGPR. Un grand nombre de méthodes, expérimentales et bioinformatiques, ont été proposées pour identifier ces gènes – nous en avions évoqué quelques unes en introduction (voir sections 1.4.1 et 1.4.2).

Ces méthodes pourraient être avantageusement combinées à AutoGPR pour d'une part tenir compte de l'essentialité des gènes candidats et d'autre part définir la place de ces nouveaux gènes dans les GPR des réactions qui leur sont associées. La Figure 55 ci-après illustre une manière d'intégrer AutoGPR avec ces approches. La combinaison de ces approches avec AutoGPR serait d'autant plus bénéfique qu'elles exploitent des données réellement complémentaires : AutoGPR ignore en effet

totalement les fonctions des gènes[80] et la plupart des méthodes de recherche de gènes candidats n'exploitent pas leur essentialité et prennent en considération leurs places dans les réseaux métaboliques de manière très simple.

13.3.2 Composantes RESEAU et BIOMASSE fixes

L'hypothèse selon laquelle les composantes RESEAU et BIOMASSE sont considérées correctes limite également la recherche de corrections par AutoGPR. Nous avions ainsi vu pour B. subtilis que trois interprétations d'incohérence mettaient en jeu des corrections GPR associées à des modifications des composantes RESEAU et BIOMASSE (gènes *fbaA, yumC* et *trxB*, voir section 12.3.2). Les essentialités de réactions étant initialement fausses et incompatibles avec les essentialités de gènes, aucune correction GPR ne pouvaient être proposée par AutoGPR.

Dans ce cas également, AutoGPR peut être avantageusement associé à des stratégies de correction des autres composantes (voir Figure 55).

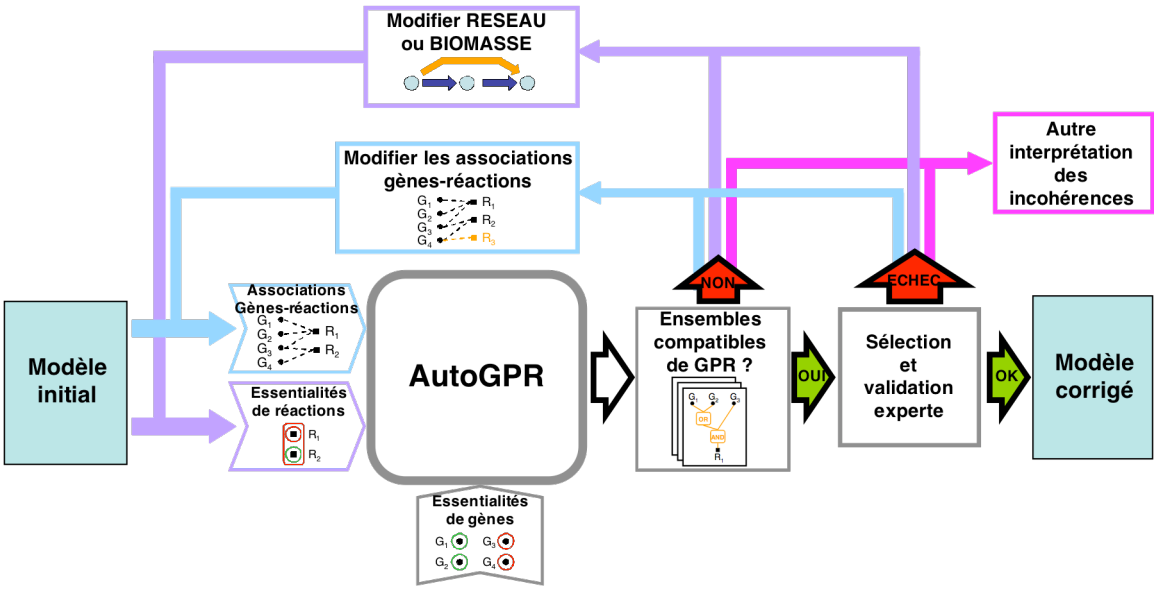


Figure 55. Intégration de stratégies de correction des liens gènes-réactions et des composantes RESEAU et BIOMASSE avec AutoGPR.

Suivant par exemple le constat qu'aucune correction GPR n'existe, ces stratégies pourraient être mises en œuvre pour corriger les composantes RESEAU ou

[80] sauf, indirectement, à travers les liens gènes-réactions prédéfinis ; mais ceux-ci ne sont pas modifiés par AutoGPR.

BIOMASSE et modifier les essentialités de réactions. AutoGPR exploiterait alors ces nouvelles essentialités afin de rechercher à nouveau des liens GPR compatibles.

Des interactions existent entre AutoGPR et ces stratégies de correction des autres composantes. En effet, si aucune correction GPR n'existe pour une incohérence de gène essentiel, cela signifie qu'il manque une ou un groupe de réactions essentielles à associer au gène. Ce constat guide la recherche de corrections dans les autres composantes : ces dernières doivent rendre essentiel au moins un groupe de réactions parmi celles associées au gène pour qu'une correction GPR compatible soit ensuite envisageable. Inversement, une incohérence de gène non-essentiel sans correction GPR « demande » aux méthodes de correction des autres composantes à rendre non-essentielle la réaction liée à ce gène.

13.3.3 GPR constantes sur tous les milieux

Enfin, l'hypothèse d'uniformité des GPR sur tous les milieux est mise à mal par la présence de régulations modifiant l'expression des gènes en fonction des milieux. Dans ces cas, AutoGPR ne peut souvent pas concilier les essentialités sur les différents milieux, ces dernières étant influencées par les régulations (voir l'exemple des régulations pour S. cerevisiae sur milieu glucose).

La méthode AutoGPR pourrait être étendue pour, dans un deuxième temps, chercher à introduire les règles de régulation les plus probables permettant d'expliquer les essentialités distinctes sur les milieux. Ces règles s'exprimant déjà sous forme booléenne dans le cadre de modélisation rFBA (Covert et al. 2001), elles pourraient bénéficier du cadre de raisonnement logique mis en place dans AutoGPR. Ce type de raisonnement a déjà été effectué manuellement pour corriger les interactions de régulation d'un modèle d'*E. coli* à l'aide de données phénotypiques (Covert et al. 2004).

13.4 Perspectives d'utilisation des délétions multiples

Bien que nous n'ayons ici appliqué AutoGPR qu'à des phénotypes de délétions simples de gènes, AutoGPR peut théoriquement prendre en compte les essentialités de n'importe quels groupes de gènes inclus dans le modèle. La disponibilité de phénotypes de délétions multiples contribuerait d'ailleurs avantageusement à l'efficacité d'AutoGPR. D'une part, ces essentialités ajouteraient des spécifications

supplémentaires aux GPR, réduisant de ce fait le nombre de propositions et, d'autre part, l'utilisation de délétions multiples permet de perturber et ainsi d'étudier un plus grand nombre de fonctions biologiques, comme constaté dans des études explorant ce sujet (Deutscher et al. 2006; Behre et al. 2007; Deutscher et al. 2008).

Toutefois, le nombre de combinaisons de délétions multiples augmente exponentiellement avec leur taille, rendant extrêmement lourde la génération exhaustive de tels ensembles de données. AutoGPR pourrait être exploité afin de proposer les délétions multiples les plus intéressantes à réaliser dans le cadre de la recherche de GPR. En effet, les grands nombres de propositions de corrections proposées par AutoGPR pourraient être significativement réduits en incluant des résultats de phénotypes de délétions multiples bien choisies. Ces dernières pourraient par exemple être sélectionnées selon un critère évaluant l'intérêt des spécifications qu'elles apporteraient.

CONCLUSIONS ET PERSPECTIVES

14 Contributions principales

Le principal objectif de cette thèse aura été de montrer que les capacités d'analyses des modèles mathématiques du métabolisme pouvaient être avantageusement mises à profit pour élucider le métabolisme des microorganismes. Plus spécifiquement, cette thèse se sera concentrée sur un type de modèles – les modèles globaux du métabolisme – et leur aptitude à exploiter une catégorie de données expérimentales auparavant difficilement interprétable à la lumière du métabolisme – les phénotypes de croissance. Elle se sera appuyée pour cela sur l'organisme *Acinetobacter baylyi* ADP1, dont nous aurons reconstruit puis corrigé le modèle métabolique à l'aide des phénotypes de croissance de ses mutants. Nous allons reprendre ici succinctement les principales conclusions de nos travaux pour en souligner nos contributions.

Dans un premier temps, nous avons abordé le problème de la reconstruction des modèles globaux à partir de la connaissance des voies métaboliques, cette connaissance étant pour la majorité des microorganismes principalement déduite de l'annotation de leurs génomes. Ces reconstructions s'effectuent en deux étapes : (1) l'identification des activités du réseau métabolique et (2) l'adaptation du réseau au formalisme mathématique. Sur l'exemple d'*A. baylyi*, nous avons proposé un processus complet de reconstruction. Nous nous sommes largement appuyés sur les outils logiciels existants pour effectuer la première étape. Notre principale contribution à cette étape aura été de souligner l'importance de combiner diverses

sources d'informations et de prendre en compte les niveaux de confiance des activités identifiées. Les points difficiles de la deuxième étape étant moins clairement maîtrisés, nous en avons décris les principaux et proposé pour une partie d'entre eux des analyses et méthodes de résolution originales: la prise en compte de la conservation de l'énergie, la génération de métabolites spécifiques et la recherche de complexes enzymatiques. Enfin, nous avons réalisé l'ensemble de ce processus de reconstruction pour *A. baylyi*, résultant en un modèle de son métabolisme complet permettant d'effectuer des prédictions quantitatives de croissance.

La deuxième partie de nos travaux a été consacrée à l'exploitation des phénotypes de croissance par les modèles globaux. Dans la lignée de travaux précédents pour quelques autres organismes, nous avons confronté phénotypes prédits et phénotypes expérimentaux d'*A. baylyi* afin d'évaluer la cohérence du modèle avec ces observations expérimentales et de proposer, le cas échéant, des corrections au modèle. Ce travail pour *A. baylyi* nous a conduit à réaliser un nombre substantiel d'amélioration au modèle initialement reconstruit et de déduire à partir des phénotypes de nouvelles informations sur son fonctionnement métabolique. Du point de vue méthodologique, nous avons introduit une distinction formelle simple entre différentes composantes des modèles métaboliques globaux – composantes GPR, RESEAU et BIOMASSE – dont le découplage permet de rechercher indépendamment des corrections. Du point de vue logiciel, nous avons participé au développement de NemoStudio et CycSim, deux interfaces web facilitant la prédiction de phénotypes de croissance, leur confrontation aux observations expérimentales et leur interprétation à la lumière des voies métaboliques.

Enfin, la dernière partie de nos travaux s'est concentrée sur la formalisation de la recherche de corrections pour la composante GPR. Nous avons ainsi participé au développement d'une méthode – AutoGPR – déduisant automatiquement l'ensemble des corrections GPR qui lèvent les incohérences des prédictions de phénotypes. Afin de montrer la pertinence d'une telle méthode, nous l'avons appliquée à la recherche de corrections pour cinq modèles métaboliques distincts et comparé ses propositions aux interprétations expertes des incohérences. Dans un dernier temps, nous avons répertorié les principales limites de cette méthode et proposé des améliorations

permettant de les surmonter et d'intégrer AutoGPR avec des stratégies de corrections des autres composantes des modèles.

15 Revue de travaux sur le même sujet effectués sur la période de la thèse (2005–2009)

Le thème de recherche de nos travaux s'est révélé être extrêmement actif durant les années de notre thèse. Des avancées sont ainsi venues progressivement compléter l'état de l'art présenté en introduction de ce manuscrit. Afin de situer nos travaux dans leur contexte actuel et d'en présenter de manière plus complète les perspectives, nous effectuerons dans cette section un rapide tour d'horizon des travaux publiés entre fin 2005 et début 2009, et liés à la reconstruction des modèles et l'exploitation des phénotypes de croissance[81].

Initialement réalisées seulement par un nombre réduit d'équipes de recherche, les reconstructions de modèles métaboliques ont rapidement gagné en popularité à partir de fin 2005, impliquant aujourd'hui plus d'une dizaine d'équipes. Le nombre de nouveaux modèles publiés chaque année a lui aussi augmenté significativement, passant de 4 modèles publiés en 2006 à une quinzaine en 2008, et déjà une dizaine pour le premier semestre de 2009[82]. Parmi ces reconstructions, une proportion toujours plus large utilise les données expérimentales de phénotypes de croissance pour évaluer la qualité du modèle et éventuellement le compléter. C'est le cas notamment des modèles d'*E. coli* (Joyce et al. 2006), *B. subtilis* (Oh et al. 2007; Henry et al. 2009), *Pseudomonas aeruginosa* (Oberhardt et al. 2008) et *Pseudomonas putida* (Puchałka et al. 2008), *Mycoplasma genitalium* (Suthers et al. 2009), *Geobacter sulfurreducens* (Segura et al. 2008), et *S. cerevisiae* (Snitkin et al. 2008).

Cette hausse du nombre de reconstructions s'est accompagnée du développement d'outils facilitant la transformation d'un réseau métabolique en un modèle mathématique. Tout d'abord, la base de données métabolique MetaCyc s'est

[81] Nous nous excusons auprès des lecteurs des quelques répétitions avec la revue sur les modèles métaboliques incluse en introduction.

[82] Un tableau disponible à l'adresse http://gcrg.ucsd.edu/In_Silico_Organisms/Other_Organisms cherche à répertorier les modèles métaboliques globaux.

progressivement adaptée aux contraintes posées par la modélisation : les compartiments cellulaires sont désormais pris en compte et des efforts de curation ont été réalisés pour équilibrer systématiquement toutes les réactions[83]. Des méthodes de reconstruction spécifiquement adaptées aux modèles à base de contraintes ont également vu le jour (Feist et al. 2009; Durot et al. 2009). Celle développée par exemple par DeJongh et al (2007) reconstruit progressivement les modèles en assemblant des « sous-modèles » des fonctions métaboliques représentées dans la base de données SEED et dont le bon fonctionnement est vérifié isolément. D'autres méthodes ont été développées pour corriger des aspects spécifiques aux modèles. Ainsi, Kumar et al (2007) ont proposé les méthodes GapFill et GapFind pour détecter et combler les impasses dans les voies métaboliques, Kümmel et al (2006b) ont introduit des règles pour établir la réversibilité des réactions, et Gevorgyan et al (2008) ont élaboré un algorithme pour détecter les incohérences de stœchiométrie entre réactions sans recourir aux formules chimiques des métabolites.

S'agissant de la prédiction des phénotypes de croissance, quelques nouvelles méthodes ont également été introduites. Kaleta et al (2008) ont par exemple exploité la notion d'*organisation chimique*[84] pour prédire les phénotypes de croissance de modèles intégrant métabolisme et régulation. Whelan & King (2008) ont pour leur part développé un modèle logique du métabolisme de *S. cerevisiae* dans le but d'exploiter des techniques d'inférence logique afin d'améliorer le modèle à partir de données de phénotypes. Plusieurs méthodes ont également été proposées pour prédire à grande échelle les environnements de croissance d'organismes à partir de leurs réseaux métaboliques ; c'est le cas des travaux de Borenstein et al (2008) et de Handorf et al (2008) à l'aide de graphes métaboliques, et d'Imielinski et al (2006) à partir des modèles à base de contraintes. Parallèlement, la recherche des ensembles essentiels minimaux de gènes a motivé plusieurs projets distincts. Klamt et al (2004) avaient ainsi développé une méthode basée sur les modes élémentaires pour

[83] Voir les améliorations apportées à MetaCyc à l'adresse suivante : http://metacyc.org/release-notes.shtml

[84] Une *organisation chimique* est un ensemble de métabolite ayant les propriétés de *clôture* (aucun métabolite extérieur à l'ensemble ne peut être produit par une réaction à partir de métabolites de l'organisation chimique) et d'*autosuffisance* (chaque métabolite consommé dans l'organisation peut être recréé à partir d'autres métabolites de l'organisation à une vitesse suffisante pour assurer sa présence).

déterminer exhaustivement les ensembles minimaux de gènes essentiels (appelés *Minimal Cut Sets*), Cette méthode et une exploration systématique étant trop complexes pour être appliquées aux modèles métaboliques d'échelle globale, Deutscher et al (2006) et Imielinski & Belta (2008) ont exploré les ensembles essentiels de gènes en développant des méthodes applicables à cette échelle, basées respectivement sur un échantillonnage des ensembles essentiels et des modes élémentaires partiels. Behre et al (2007) ont quand à eux introduit une mesure quantitative de la robustesse aux délétions multiples permettant d'évaluer de manière plus complète[85] la robustesse d'un réseau aux perturbations génétiques. Enfin, Deutscher et (2008) ont élaboré un indicateur basé sur la théorie des jeux permettant d'exploiter les phénotypes de perturbations génétiques multiples pour quantifier la contribution d'un gène à la réalisation de fonction biologiques.

Dans le même esprit qu'AutoGPR ont été proposées des méthodes de corrections automatiques des modèles permettant de résoudre les prédictions incohérentes de phénotypes de croissances. Reed, Patel et al (2006) ont ainsi développé un algorithme recherchant le nombre minimal de réactions à ajouter à un modèle pour lui permettre de prédire la croissance sur un environnement particulier. Kumar & Maranas (2009) ont quant à eux élaboré la méthode GrowMatch qui vise à réconcilier prédictions et observations d'essentialités de gènes en modifiant principalement la composante RESEAU des modèles (ajout/suppression de réactions, changement de leurs réversibilités). Cette dernière méthode complèterait de manière appropriée AutoGPR pour élaborer des corrections plus complètes des modèles.

Enfin, il est intéressant de signaler une initiative visant à automatiser entièrement la recherche des gènes des activités orphelines par des approches génétiques (King et al. 2009). En se basant sur le modèle logique du métabolisme de *S. cerevisiae* mentionné plus haut (Whelan & King 2008), King et al ont implémenté une méthode qui, pour chaque activité orpheline, (1) sélectionne des gènes candidats, (2) infère les expériences de génétique (délétions de gènes, environnement) à réaliser pour identifier le bon candidat, (3) effectue automatiquement les expériences à l'aide d'un robot, et (4) conclut à partir des phénotypes observés expérimentalement. Ce

[85] par rapport aux délétions uniquement simples.

processus, appliqué à la recherche des gènes de 13 activités orphelines, a permis d'identifier 20 candidats dont les fonctions ont été confirmées pour une partie d'entre eux par des tests biochimiques directs ou des recherches dans la littérature.

16 Perspectives

Nous conclurons ce manuscrit en évoquant quelques perspectives ouvertes par l'amélioration de la reconstruction des modèles et de leur capacité à intégrer des données expérimentales.

Il semble en effet que la reconstruction des réseaux métaboliques devienne une extension naturelle de l'annotation des génomes. L'essor des outils de reconstruction et des bases de données métaboliques (notamment BioCyc et KEGG) témoignent de cette tendance. De même, les plateformes d'annotation actuelles[86] évoluent progressivement afin de replacer les fonctions des gènes dans le contexte de processus biologiques complets – à l'instar des sous-systèmes définis dans SEED (Overbeek et al. 2005). Ces plateformes reconstruisent désormais systématiquement les réseaux métaboliques correspondant à chaque génome (souvent à l'aide de KEGG ou BioCyc) et cherchent à en exploiter la vision par voies métaboliques pour préciser les fonctions des gènes et compléter les annotations. Ces plateformes se limitent pour le moment à capturer les annotations expertes de la fonction des gènes ; il est probable que ces outils évolueront pour prendre en compte des informations expertes sur les voies métaboliques elles-mêmes, comme cela est déjà proposé aux curateurs des bases de données BioCyc (Caspi et al. 2008).

Les modèles métaboliques peuvent également prétendre à devenir des compléments systématiques à l'annotation des génomes. En effet, les méthodes introduites dans cette thèse et, plus largement, les travaux portant sur la reconstruction des modèles vont vraisemblablement réussir à lever les contraintes liées au formalisme mathématique, qui entravent actuellement la création d'un modèle à partir d'un réseau métabolique. De plus, comme suggéré dans cette thèse pour les phénotypes de croissance, l'aptitude des modèles à exploiter des données

[86] Notamment MaGe (Vallenet et al. 2006), IMG (Markowitz et al. 2009) ou SEED (Aziz et al. 2008).

expérimentales de diverses natures va très certainement promouvoir leur utilisation pour élucider le métabolisme des organismes et compléter l'annotation de leurs génomes. Cet argumentaire est à la base de Microme, un projet européen devant débuter fin 2009 dans lequel le Genoscope est impliqué. L'objectif de Microme sera de fournir des méthodes et des infrastructures logicielles pour reconstruire automatiquement et effectuer la curation experte des réseaux et modèles métaboliques d'organismes procaryotes. Microme intègrera notamment des méthodes de confrontations des modèles aux données expérimentales qui suggèreront des pistes de curation aux experts.

La mise en place d'infrastructures du type envisagé par Microme ouvre alors la voie à la reconstruction en grand nombre de modèles métaboliques, offrant de nouvelles perspectives à la communauté scientifique.

Tout d'abord, la disponibilité de modèles métaboliques pour un grand nombre d'organismes distribués sur l'arbre de la vie fournirait de nouveaux outils pour étudier l'évolution des organismes, notamment procaryotes. Les modèles relient en effet directement les gènes à leurs rôles dans le métabolisme, permettant d'étudier en retour les contraintes posées par le métabolisme sur l'évolution des gènes. Quelques travaux ont déjà été réalisés dans cet esprit sur un nombre limité d'organismes (Pál et al. 2005; Pál et al. 2006), mais l'utilisation de modèles en plus grand nombre permettra très certainement d'élargir leur champ d'applications dans ce domaine.

Ensuite, des modèles systématiquement reconstruits sont autant d'outils à la disposition de la communauté scientifique pour interpréter les grands ensembles de données expérimentales. Les progrès techniques offrent aux expérimentateurs la possibilité de mesurer à grande échelle une large variété de grandeurs liées aux entités biologiques : par exemple les concentrations métaboliques, flux de réactions métaboliques, expressions de gènes, concentrations de protéines et, au niveau macroscopique, phénotypes de croissance. Cependant, il s'avère que ces ensembles de données ne peuvent prendre tout leurs sens que lorsqu'ils sont interprétés à la lumière du fonctionnement biochimique réel de la cellule. Les modèles permettent justement d'intégrer ces données et de les mettre en regard du fonctionnement métabolique modélisé. En disposant de modèles métaboliques pour un grand nombre d'organismes, les logiciels implémentant les méthodes d'intégration de données, telle

que CycSim pour les phénotypes de croissance, constitueraient des outils bienvenus pour interpréter ces données.

Enfin, l'ingénierie du métabolisme est un autre domaine susceptible de bénéficier de la disponibilité de nombreux modèles. Ces ensembles de modèles constitueraient en effet de véritables répertoires virtuels d'organismes et de voies métaboliques, permettant de réaliser et de tester *in silico* un grand nombre de « constructions métaboliques », avant même toute expérience en laboratoire. Ils pourraient notamment être employés pour (1) sélectionner les organismes aux caractéristiques métaboliques les plus adaptés à l'objectif métabolique, (2) prédire les performances théoriques de modifications métaboliques envisagées ou (3) suggérer automatiquement des modifications métaboliques répondant à un objectif, constituant de ce fait une véritable boîte à outils numérique à la disposition des ingénieurs du métabolisme.

REFERENCES BIBLIOGRAPHIQUES

Abbott, A., 2005. Medics braced for fresh superbug. *Nature*, 436(7052), 758.

Abbott, B.J., Laskin, A.I. & McCoy, C.J., 1974. Effect of growth rate and nutrient limitation on the composition and biomass yield of Acinetobacter calcoaceticus. *Appl Microbiol*, 28(1), 58–63.

Abd-El-Haleem, D., 2003. Acinetobacter: environmental and biotechnological applications. *Afr. J. Biotechnol.*, 2(4), 71–74.

Aghaie, A., Lechaplais, C., Sirven, P., Tricot, S., Besnard-Gonnet, M., Muselet, D., de Berardinis, V., Kreimeyer, A., Gyapay, G., Salanoubat, M. & Perret, A., 2008. New insights into the alternative d-glucarate degradation pathway. *J Biol Chem*, 283(23), 15638–15646.

Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N. & Mekalanos, J.J., 2002. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *Proc Natl Acad Sci U S A*, 99(2), 966–971.

Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C.M., Podowski, R.M., Naslund, A.K., Eriksson, A., Winkler, H.H. & Kurland, C.G., 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, 396(6707), 133-140.

Arigoni, F., Talabot, F., Peitsch, M., Edgerton, M.D., Meldrum, E., Allet, E., Fish, R., Jamotte, T., Curchod, M.L. & Loferer, H., 1998. A genome-based approach for the identification of essential bacterial genes. *Nature Biotechnology*, 16(9), 851-6.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V.,

Wilke, A. & Zagnitko, O., 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9, 75.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. & Mori, H., 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2, 2006.0008.

Bairoch, A., 2000. The ENZYME database in 2000. *Nucleic Acids Res*, 28(1), 304–305.

Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P., Ornston, L.N., Weissenbach, J., Marlière, P., Cohen, G.N. & Médigue, C., 2004. Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res*, 32(19), 5766–5779.

Barkai, N. & Leibler, S., 1997. Robustness in simple biochemical networks. *Nature*, 387(6636), 913-7.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I. & Schomburg, D., 2007. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res*, 35(Database issue), D511–D514.

Batada, N.N., Hurst, L.D. & Tyers, M., 2006. Evolutionary and physiological importance of hub proteins. *PLoS Computational Biology*, 2(7), e88.

Beard, D.A., Babson, E., Curtis, E. & Qian, H., 2004. Thermodynamic constraints for biochemical networks. *J Theor Biol*, 228(3), 327–333.

Beard, D.A., Liang, S. & Qian, H., 2002. Energy balance for analysis of complex metabolic networks. *Biophys J*, 83(1), 79–86.

Becker, S.A., Feist, A.M., Mo, M.L., Hannum, G., Palsson, B.Ø. & Herrgard, M.J., 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*, 2(3), 727–738.

Behre, J., Wilhelm, T., von Kamp, A., Ruppin, E. & Schuster, S., 2007. Structural robustness of metabolic networks with respect to multiple knockouts. *J Theor Biol*, 252(3), 433–441.

de Berardinis, V., Vallenet, D., Castelli, V., Besnard, M., Pinet, A., Cruaud, C., Samair, S., Lechaplais, C., Gyapay, G., Richez, C., Durot, M., Kreimeyer, A., Le Fèvre, F., Schächter, V., Pezo, V., Döring, V., Scarpelli, C., Médigue, C., Cohen, G.N., Marlière, P., Salanoubat, M. & Weissenbach, J., 2008. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Mol Syst Biol*, 4, 174.

Bergogne-Bérézin, E. & Towner, K.J., 1996. Acinetobacter spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clin Microbiol Rev*, 9(2), 148–165.

Bochner, B.R., 2009. Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews*, 33(1), 191-205.

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C., Longabaugh, W., Vuthoori, M., Whitehead, K., Madar, A., Suzuki, L., Mori, T., Chang, D., Diruggiero, J., Johnson, C.H., Hood, L. & Baliga, N.S., 2007. A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131(7), 1354-65.

Borenstein, E., Kupiec, M., Feldman, M.W. & Ruppin, E., 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc Natl Acad Sci U S A*, 105(38), 14482–14487.

Borodina, I., Krabben, P. & Nielsen, J., 2005. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res*, 15(6), 820–829.

Boyd, S. & Vandenberghe, L., 2004. *Convex Optimization*, Cambridge University Press . Available at: http://www.stanford.edu/~boyd/cvxbook/.

Breitling, R., Vitkup, D. & Barrett, M.P., 2008. New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol*, 6(2), 156–161.

Briggs, G.E. & Haldane, J.B., 1925. A Note on the Kinetics of Enzyme Action. *The Biochemical Journal*, 19(2), 338-9.

Bryan, B.A., Linhardt, R.J. & Daniels, L., 1986. Variation in composition and yield of exopolysaccharides produced by Klebsiella sp. strain K32 and Acinetobacter calcoaceticus BD4. *Appl Environ Microbiol*, 51(6), 1304–1308.

Burgard, A.P. & Maranas, C.D., 2003. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng*, 82(6), 670–677.

Burgard, A.P., Nikolaev, E.V., Schilling, C.H. & Maranas, C.D., 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*, 14(2), 301–312.

Butland, G., Babu, M., Díaz-Mejía, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O., Mori, H., Wanner, B.L., Lo, H., Wasniewski, J., Christopolous, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., Caron, S., Choi, J., Laigle, L., Nazarians-Armavil, A., Deshpande, A., Joe, S., Datsenko, K.A., Yamamoto, N., Andrews, B.J., Boone, C., Ding, H., Sheikh, B., Moreno-Hagelseib, G., Greenblatt, J.F. & Emili, A., 2008. eSGA: E. coli synthetic genetic array analysis. *Nature Methods*, 5(9), 789-95.

Carpenter, A.E. & Sabatini, D.M., 2004. Systematic genome-wide screens of gene function. *Nat Rev Genet*, 5(1), 11–22.

Carr, E.L., Kämpfer, P., Patel, B.K.C., Gürtler, V. & Seviour, R.J., 2003. Seven novel species of Acinetobacter isolated from activated sludge. *International Journal of Systematic and Evolutionary Microbiology*, 53(Pt 4), 953-63.

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P. & Karp, P.D., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 36(Database issue), D623-31.

Chalker, A.F. & Lunsford, R.D., 2002. Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacology & Therapeutics*, 95(1), 1-20.

Chen, T., Siu, L., Lee, Y., Chen, C., Huang, L., Wu, R.C., Cho, W. & Fung, C., 2008. Acinetobacter baylyi as a pathogen for opportunistic infection. *Journal of Clinical Microbiology*, 46(9), 2938-44.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D. & Cherry, J.M., 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Research*, 32(Database issue), D311-314.

Cornish-Bowden, A., 2004. *Fundamentals of Enzyme Kinetics* 3 éd., London: Portland Press.

Covert, M.W., Schilling, C.H. & Palsson, B., 2001. Regulation of gene expression in flux balance models of metabolism. *J Theor Biol*, 213(1), 73–88.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O., 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987), 92–96.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. & Ashburner, M., 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database issue), D344-350.

DeJongh, M., Formsma, K., Boillot, P., Gould, J., Rycenga, M. & Best, A., 2007. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics*, 8, 139.

Deutscher, D., Meilijson, I., Kupiec, M. & Ruppin, E., 2006. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet*, 38(9), 993–998.

Deutscher, D., Meilijson, I., Schuster, S. & Ruppin, E., 2008. Can single knockouts accurately single out gene functions? *BMC Syst Biol*, 2(1), 50.

Di Ventura, B., Lemerle, C., Michalodimitrakis, K. & Serrano, L., 2006. From in vivo to in silico biology and back. *Nature*, 443(7111), 527–533.

Dole, M., 1965. The Natural History of Oxygen. *The Journal of General Physiology*, 49, 5–27.

Doten, R.C., Ngai, K.L., Mitchell, D.J. & Ornston, L.N., 1987. Cloning and genetic organization of the pca gene cluster from Acinetobacter calcoaceticus. *Journal of Bacteriology*, 169(7), 3168-3174.

Duarte, N.C., Herrgard, M.J. & Palsson, B.O., 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7), 1298–1309.

Duarte, N.C., Palsson, B.O. & Fu, P., 2004. Integrated analysis of metabolic phenotypes in Saccharomyces cerevisiae. *BMC Genomics*, 5(1), 63.

Dunn, W.B., Bailey, N.J.C. & Johnson, H.E., 2005. Measuring the metabolome: current analytical technologies. *Analyst*, 130(5), 606–625.

Durot, M., Bourguignon, P. & Schachter, V., 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1), 164-90.

Durot, M., Le Fèvre, F., de Berardinis, V., Kreimeyer, A., Vallenet, D., Combe, C., Smidtas, S., Salanoubat, M., Weissenbach, J. & Schachter, V., 2008. Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Systems Biology*, 2, 85.

Dykhuizen, D.E., Dean, A.M. & Hartl, D.L., 1987. Metabolic flux and fitness. *Genetics*, 115(1), 25–31.

Dykxhoorn, D.M., Novina, C.D. & Sharp, P.A., 2003. Killing the messenger: short RNAs that silence gene expression. *Nature Reviews. Molecular Cell Biology*, 4(6), 457-67.

Edwards, J.S., Ibarra, R.U. & Palsson, B.O., 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2), 125–130.

Edwards, J.S. & Palsson, B.O., 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, 97(10), 5528–5533.

Ellis, L.B.M., Roe, D. & Wackett, L.P., 2006. The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res*, 34(Database issue), D517–D521.

Engdahl, H.M., Hjalt, T.A. & Wagner, E.G., 1997. A two unit antisense RNA cassette test system for silencing of target genes. *Nucleic Acids Research*, 25(16), 3218-27.

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M.J.O. & Dennis, E.A., 2009. Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of Lipid Research*, 50(Supplement), S9-14.

Famili, I., Forster, J., Nielsen, J. & Palsson, B.O., 2003. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A*, 100(23), 13134–13139.

Fang, G., Rocha, E. & Danchin, A., 2005. How essential are nonessential genes? *Mol Biol Evol*, 22(11), 2147–2156.

Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V. & Palsson, B.Ø., 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3, 121.

Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L. & Palsson, B.Ø., 2009. Reconstruction of biochemical networks in microorganisms. *Nature Reviews. Microbiology*, 7(2), 129-43.

Fell, D.A., 1992. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J*, 286 (Pt 2), 313–330.

de Figueiredo, L.F., Schuster, S., Kaleta, C. & Fell, D.A., 2009. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics (Oxford, England)*, 25(1), 152-158.

Fisher, J. & Henzinger, T.A., 2007. Executable cell biology. *Nat Biotechnol*, 25(11), 1239–1249.

Forsyth, R.A., Haselbeck, R.J., Ohlsen, K.L., Yamamoto, R.T., Xu, H., Trawick, J.D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J.M., C, K.G., King, P., McCarthy, M., Malone, C., Misiner, B., Robbins, D., Tan, Z., Zhu Zy, Z., Carr, G., Mosca, D.A., Zamudio, C., Foulkes, J.G. & Zyskind, J.W., 2002. A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Molecular Microbiology*, 43(6), 1387-400.

Fournier, P., Vallenet, D., Barbe, V., Audic, S., Ogata, H., Poirel, L., Richet, H., Robert, C., Mangenot, S., Abergel, C., Nordmann, P., Weissenbach, J., Raoult, D. & Claverie, J., 2006. Comparative genomics of multidrug resistance in Acinetobacter baumannii. *PLoS Genet*, 2(1), e7.

French, C.T., Lao, P., Loraine, A.E., Matthews, B.T., Yu, H. & Dybvig, K., 2008. Large-scale transposon mutagenesis of Mycoplasma pulmonis. *Molecular Microbiology*, 69(1), 67-76.

Funahashi, A., Morohashi, M., Kitano, H. & Tanimura, N., 2003. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, 1(5), 159-162.

Gallagher, L.A., Ramage, E., Jacobs, M.A., Kaul, R., Brittnacher, M. & Manoil, C., 2007. A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3), 1009-14.

Gennis, R.B. & Stewart, V., 1996. Respiration. Dans F. C. Neidhardt, éd. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 217–261.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balázsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M.V., Grechkin, Y., Mseeh, F., Fonstein, M.Y., Overbeek, R., Barabási, A., Oltvai, Z.N. & Osterman, A.L., 2003. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol*, 185(19), 5673–5684.

Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R. & Osterman, A., 2006. Essential genes on metabolic maps. *Curr Opin Biotechnol*, 17(5), 448–456.

Gerischer, U., Jerg, B. & Fischer, R., 2008. Spotlight on the Acinetobacter baylyi beta-ketoadipate pathway: multiple levels of regulation. Dans *Acinetobacter Molecular Biology*. Norfolk, UK: Caister Academic Press, pp. 203-230.

Gevorgyan, A., Poolman, M.G. & Fell, D.A., 2008. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 24(19), 2245–2251.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kötter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D.,

Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. & Johnston, M., 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896), 387–391.

Gillespie, D.T., 2007. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58, 35–55.

Glasner, J.D., Liss, P., Plunkett, G., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F.R. & Perna, N.T., 2003. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Research*, 31(1), 147-151.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O. & Venter, J.C., 2006. Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A*, 103(2), 425–430.

Gong, X., Fan, S., Bilderbeck, A., Li, M., Pang, H. & Tao, S., 2008. Comparative analysis of essential genes and nonessential genes in Escherichia coli K12. *Molecular Genetics and Genomics*, 279(1), 87-94.

Gutnick, D.L. & Bach, H., 2008. Potential Application of Acinetobacter in Biotechnology. Dans U. Gerischer, éd. *Acinetobacter Molecular Biology*. Norfolk, UK: Caister Academic Press, pp. 231–264.

Hahn, M.W. & Kern, A.D., 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4), 803-6.

Handorf, T., Christian, N., Ebenhöh, O. & Kahn, D., 2008. An environmental perspective on metabolism. *Journal of Theoretical Biology*, 252(3), 530-7.

Handorf, T., Ebenhöh, O. & Heinrich, R., 2005. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *Journal of Molecular Evolution*, 61(4), 498-512.

Hare, R.S., Walker, S.S., Dorman, T.E., Greene, J.R., Guzman, L.M., Kenney, T.J., Sulavik, M.C., Baradaran, K., Houseweart, C., Yu, H., Foldes, Z., Motzer, A., Walbridge, M., Shimer, G.H. & Shaw, K.J., 2001. Genetic footprinting in bacteria. *Journal of Bacteriology*, 183(5), 1694-706.

Harrison, R., Papp, B., Pál, C., Oliver, S.G. & Delneri, D., 2007. Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A*, 104(7), 2307–2312.

Hayes, F., 2003. Transposon-based strategies for microbial functional genomics and proteomics. *Annual Review of Genetics*, 37, 3-29.

Heinrich, R. & Rapoport, T.A., 1974. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *European Journal of Biochemistry / FEBS*, 42(1), 89-95.

Henry, C., Zinner, J., Cohoon, M. & Stevens, R., 2009. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biology*, 10(6), R69.

Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V., 2007. Thermodynamics-based Metabolic Flux Analysis. *Biophys J*, 92(5), 1792–1805.

Hofestädt, R., 2003. Petri nets and the simulation of metabolic networks. *In Silico Biology*, 3(3), 321-2.

Hucka, M., Finney, A., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C. & Kitano, H., 2004. Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Syst Biol (Stevenage)*, 1(1), 41–53.

Hunter, P.J. & Borg, T.K., 2003. Integration from proteins to organs: the Physiome Project. *Nature Reviews. Molecular Cell Biology*, 4(3), 237-43.

Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O. & Venter, J.C., 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. *Science (New York, N.Y.)*, 286(5447), 2165-9.

Ibarra, R.U., Edwards, J.S. & Palsson, B.O., 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912), 186–189.

Imielinski, M. & Belta, C., 2008. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Systems Biology*, 2(1), 40.

Imielinski, M., Belta, C., Halasz, A. & Rubin, H., 2005. Investigating metabolite essentiality through genome-scale analysis of Escherichia coli production capabilities. *Bioinformatics*, 21(9), 2008–2016.

Imielinski, M., Belta, C., Rubin, H. & Halász, A., 2006. Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. *Biophys J*, 90(8), 2659–2672.

Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., Chun-Rong, L., Guenthner, D., Bovee, D., Olson, M.V. & Manoil, C., 2003. Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A*, 100(24), 14339–14344.

Janssen, D.B., Dinkla, I.J.T., Poelarends, G.J. & Terpstra, P., 2005. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol*, 7(12), 1868–1882.

Jeong, H., Mason, S.P., Barabási, A.L. & Oltvai, Z.N., 2001. Lethality and centrality in protein networks. *Nature*, 411(6833), 41-2.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L., 2000. The large-scale organization of metabolic networks. *Nature*, 407(6804), 651–654.

Ji, Y., Zhang, B., Van, S.F., Horn, Warren, P., Woodnutt, G., Burnham, M.K. & Rosenberg, M., 2001. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, 293(5538), 2266-9.

Joyce, A.R. & Palsson, B.Ø., 2006. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3), 198–210.

Joyce, A.R., Reed, J.L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S.A., Palsson, B.Ø. & Agarwalla, S., 2006. Experimental and computational assessment of conditionally essential genes in Escherichia coli. *J Bacteriol*, 188(23), 8259–8271.

Juni, E. & Janik, A., 1969. Transformation of Acinetobacter calco-aceticus (Bacterium anitratum). *Journal of Bacteriology*, 98(1), 281-8.

Juni, E., 1972. Interspecies transformation of Acinetobacter: genetic evidence for a ubiquitous genus. *J Bacteriol*, 112(2), 917–931.

Kacser, H. & Burns, J.A., 1973. The control of flux. *Symposia of the Society for Experimental Biology*, 27, 65-104.

Kaleta, C., Centler, F., Fenizio, P.S.D. & Dittrich, P., 2008. Phenotype prediction in regulated metabolic networks. *BMC Syst Biol*, 2(1), 37.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y., 2007. KEGG for linking genomes to life and the environment. *Nucl. Acids Res.*, 36, D480–D484.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue), D354–D357.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M., 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue), D277–D280.

Kang, Y., Durfee, T., Glasner, J.D., Qiu, Y., Frisch, D., Winterberg, K.M. & Blattner, F.R., 2004. Systematic mutagenesis of the Escherichia coli genome. *J Bacteriol*, 186(15), 4921–4930.

Karp, P.D., Paley, S. & Romero, P., 2002. The Pathway Tools software. *Bioinformatics*, 18 Suppl 1, S225–S232.

Keseler, I.M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G. & Karp, P.D., 2009. EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Research*, 37(Database issue), D464-70.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B. & Oliver, S.G., 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971), 247–252.

King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E. & Clare, A., 2009. The Automation of Science. *Science*, 324(5923), 85-89.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H. & Mori, H., 2005. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res*, 12(5), 291–299.

Kitano, H., 2002. Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560), 1662-4.

Kitano, H., 2007. Towards a theory of biological robustness. *Mol Syst Biol*, 3, 137.

Klamt, S. & Gilles, E.D., 2004. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2), 226–234.

Klamt, S., Saez-Rodriguez, J. & Gilles, E.D., 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol*, 1, 2.

Klipp, E., Heinrich, R. & Holzhütter, H., 2002. Prediction of temporal gene expression. Metabolic opimization by re-distribution of enzyme activities. *Eur J Biochem*, 269(22), 5406–5413.

Knoll, A.H., 2003. The geological consequences of evolution. *Geobiology*, 1(1), 3-14.

Knuth, K., Niesalla, H., Hueck, C.J. & Fuchs, T.M., 2004. Large-scale identification of essential Salmonella genes by trapping lethal insertions. *Molecular Microbiology*, 51(6), 1729-44.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C.,

Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Débarbouille, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Coq, D.L., Masson, A., Mauël, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F.M.L., Sekiguchi, J., Sekowska, A., Séror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., Dijl, J.M.V., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U. & Ogasawara, N., 2003. Essential Bacillus subtilis genes. *Proc Natl Acad Sci U S A*, 100(8), 4678–4683.

Koch, I., Junker, B.H. & Heiner, M., 2005. Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics (Oxford, England)*, 21(7), 1219-26.

Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews. Microbiology*, 1(2), 127-36.

Koshland, D.E., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2), 98-104.

Kuepfer, L., Sauer, U. & Blank, L.M., 2005. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res*, 15(10), 1421–1430.

Kumar, V.S., Dasika, M.S. & Maranas, C.D., 2007. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8, 212.

Kumar, V.S. & Maranas, C.D., 2009. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Computational Biology*, 5(3), e1000308.

Kümmel, A., Panke, S. & Heinemann, M., 2006a. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol*, 2, 2006.0034.

Kümmel, A., Panke, S. & Heinemann, M., 2006b. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7, 512.

Le Fèvre, F., Smidtas, S., Combe, C., Durot, M., d'Alché-Buc, F. & Schachter, V., 2009. CycSim - an online tool for exploring and experimenting with genome-

scale metabolic models. *Bioinformatics (Oxford, England)*, 25(15), 1987-1988.

Le Fèvre, F., Smidtas, S. & Schächter, V., 2007. Cyclone: java-based querying and computing with Pathway/Genome databases. *Bioinformatics*, 23(10), 1299–1300.

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L. & Hucka, M., 2006. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*, 34(Database issue), D689–D691.

Lee, J.M., Gianchandani, E.P., Eddy, J.A. & Papin, J.A., 2008. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol*, 4(5), e1000086.

Lemerle, C., Di Ventura, B. & Serrano, L., 2005. Space as the final frontier in stochastic simulations of biological systems. *FEBS Letters*, 579(8), 1789-94.

Lespinet, O. & Labedan, B., 2006a. ORENZA: a web resource for studying ORphan ENZyme activities. *BMC Bioinformatics*, 7, 436.

Lespinet, O. & Labedan, B., 2006b. Orphan enzymes could be an unexplored reservoir of new drug targets. *Drug Discovery Today*, 11(7-8), 300-5.

Liberati, N.T., Urbach, J.M., Miyata, S., Lee, D.G., Drenkard, E., Wu, G., Villanueva, J., Wei, T. & Ausubel, F.M., 2006. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proc Natl Acad Sci U S A*, 103(8), 2833–2838.

Liebermeister, W. & Klipp, E., 2006. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor Biol Med Model*, 3, 41.

Löfberg, J., 2004. YALMIP : A Toolbox for Modeling and Optimization in MATLAB. Dans *Proceedings of the CACSD Conference*. Taipei, Taiwan. Available at: http://control.ee.ethz.ch/~joloef/yalmip.php.

Ma, H. & Zeng, A., 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2), 270–277.

Mahadevan, R. & Schilling, C.H., 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4), 264–276.

Makula, R.A., Lockwood, P.J. & Finnerty, W.R., 1975. Comparative analysis of the lipids of Acinetobacter species grown on hexadecane. *J Bacteriol*, 121(1), 250–258.

Markowitz, V.M., Mavromatis, K., Ivanova, N.N., Chen, I.A., Chu, K. & Kyrpides, N.C., 2009. IMG ER: A System for Microbial Genome Annotation Expert Review and Curation. *Bioinformatics (Oxford, England)*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19561336 [Accédé Juillet 29, 2009].

Maskow, T. & von Stockar, U., 2005. How reliable are thermodynamic feasibility statements of biochemical pathways? *Biotechnol Bioeng*, 92(2), 223–230.

May, R.M., 2004. Uses and abuses of mathematics in biology. *Science (New York, N.Y.)*, 303(5659), 790-3.

McGovern, P.E., Glusker, D.L., Exner, L.J. & Voigt, M.M., 1996. Neolithic resinated wine. *Nature*, 381(6582), 480-481.

van der Meer, J.R., de Vos, W.M., Harayama, S. & Zehnder, A.J., 1992. Molecular mechanisms of genetic adaptation to xenobiotic compounds. *Microbiological Reviews*, 56(4), 677-94.

Metzgar, D., Bacher, J.M., Pezo, V., Reader, J., Döring, V., Schimmel, P., Marlière, P. & de Crécy-Lagard, V., 2004. Acinetobacter sp. ADP1: an ideal model organism for genetic analysis and genome engineering. *Nucleic Acids Res*, 32(19), 5780–5790.

Médigue, C. & Moszer, I., 2007. Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol*, 158(10), 724–736.

Michaelis, L. & Menten, M.L., 1913. Die Kinetik der Invertinwirkung. *Biochem. Z*, 49(333), 148.

Mitchell, A., Romano, G.H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O. & Pilpel, Y., 2009. Adaptive prediction of environmental changes by microorganisms. *Nature*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19536156 [Accédé Juillet 8, 2009].

Moisdon, J., 2000. *Recherche opérationnelle. Programmation linéaire*, Ecole des Mines de Paris.

Moraru, I.I., Schaff, J.C., Slepchenko, B.M., Blinov, M.L., Morgan, F., Lakshminarayana, A., Gao, F., Li, Y. & Loew, L.M., 2008. Virtual Cell modelling and simulation software environment. *IET Systems Biology*, 2(5), 352-62.

Motter, A.E., Gulbahce, N., Almaas, E. & Barabási, A., 2008. Predicting synthetic rescues in metabolic networks. *Mol Syst Biol*, 4, 168.

Murphy, K.C., Campellone, K.G. & Poteete, A.R., 2000. PCR-mediated gene replacement in Escherichia coli. *Gene*, 246(1-2), 321–330.

Mushegian, A.R. & Koonin, E.V., 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10268-73.

Neidhardt, F.C., 1996. The Enteric Bacterial Cell and the Age of Bacteria. Dans F. C. Neidhardt, éd. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 1-4.

Neidhardt, F.C. & Umbarger, H.E., 1996. Chemical composition of Escherichia coli. Dans F. C. Neidhardt, éd. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 13-16.

Neidhardt, F.C. éd., 1996. *Escherichia coli and Salmonella: cellular and molecular biology* 2 éd., Washington, D.C.: ASM Press.

Neijssel, O.M., Teixeira de Mattos, M.J. & Tempest, D.W., 1996. Growth Yield and Energy Distribution. Dans F. C. Neidhardt, éd. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 1683-1692.

Noble, D., 2002. Modeling the heart--from genes to cells to the whole organ. *Science*, 295(5560), 1678-82.

Oberhardt, M.A., Puchałka, J., Fryer, K.E., Santos, V.A.P.M.D. & Papin, J.A., 2008. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J Bacteriol*, 190(8), 2790–2803.

Oh, Y., Palsson, B.O., Park, S.M., Schilling, C.H. & Mahadevan, R., 2007. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem*, 282(39), 28791–28799.

Oliveira, A.P., Nielsen, J. & Förster, J., 2005. Modeling Lactococcus lactis using a genome-scale flux model. *BMC Microbiology*, 5, 39.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17), 5691–5702.

Paley, S.M. & Karp, P.D., 2006. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res*, 34(13), 3771–3778.

Palmen, R. & Hellingwerf, K.J., 1997. Uptake and processing of DNA by Acinetobacter calcoaceticus--a review. *Gene*, 192(1), 179-190.

Papp, B., Pál, C. & Hurst, L.D., 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992), 661–664.

Park, J.H., Lee, S.Y., Kim, T.Y. & Kim, H.U., 2008. Application of systems biology for bioprocess development. *Trends Biotechnol*, 26(8), 404–412.

Pál, C., Papp, B. & Lercher, M.J., 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12), 1372–1375.

Pál, C., Papp, B., Lercher, M.J., Csermely, P., Oliver, S.G. & Hurst, L.D., 2006. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084), 667–670.

Pouliot, Y. & Karp, P.D., 2007. A survey of orphan enzyme activities. *BMC Bioinformatics*, 8, 244.

du Preez, J.C., Lategan, P.M. & Toerien, D.F., 1984. Influence of the growth rate on the macromolecular composition of A cinetobacter calcoaceticus in carbon-limited chemostat culture. *FEMS Microbiology Letters*, 23, 71–75.

Price, N.D., Reed, J.L. & Palsson, B.O., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11), 886–897.

Puchałka, J., Oberhardt, M.A., Godinho, M., Bielecka, A., Regenhardt, D., Timmis, K.N., Papin, J.A. & Santos, V.A.P.M.D., 2008. Genome-Scale Reconstruction and Analysis of the Pseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. *PLoS Comput Biol*, 4(10), e1000210.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabási, A.L., 2002. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551–1555.

Raymond, J. & Segrè, D., 2006. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311(5768), 1764–1767.

Reams, A.B. & Neidle, E.L., 2004. Selection for gene clustering by tandem duplication. *Annual Review of Microbiology*, 58, 119-42.

Reddy, V.N., Liebman, M.N. & Mavrovouniotis, M.L., 1996. Qualitative analysis of biochemical reaction systems. *Computers in Biology and Medicine*, 26(1), 9-24.

Reed, J.L., Famili, I., Thiele, I. & Palsson, B.O., 2006. Towards multidimensional genome annotation. *Nat Rev Genet*, 7(2), 130–141.

Reed, J.L., Patel, T.R., Chen, K.H., Joyce, A.R., Applebee, M.K., Herring, C.D., Bui, O.T., Knight, E.M., Fong, S.S. & Palsson, B.O., 2006. Systems approach to

refining genome annotation. *Proc Natl Acad Sci U S A*, 103(46), 17480–17484.

Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O., 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9), R54.

Reich, K.A., Chovan, L. & Hessler, P., 1999. Genome scanning in Haemophilus influenzae for identification of essential genes. *Journal of Bacteriology*, 181(16), 4961-8.

Reitzer, L.J., 1996. Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine, and D-alanine. Dans F. C. Neidhardt, éd. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 391–407.

Ren, Q., Kang, K.H. & Paulsen, I.T., 2004. TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res*, 32(Database issue), D284–D288.

Reznikoff, W.L. & Winterberg, K.M., 2008. Transposon-based strategies for the identification of essential bacterial genes. Dans A. L. Osterman & S. Y. Gerdes, éd. *Microbial Gene Essentiality: Protocols and bioinformatics*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 13-26.

Ro, D., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y., Withers, S.T., Shiba, Y., Sarpong, R. & Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086), 940–943.

Rocha, E.P.C. & Danchin, A., 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet*, 34(4), 377–378.

Rockafellar, R., 1970. *Convex Analysis*, Princeton University Press.

Romero, P.R. & Karp, P., 2001. Nutrient-related analysis of pathway/genome databases. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 471-82.

Saghatelian, A., Trauger, S.A., Want, E.J., Hawkins, E.G., Siuzdak, G. & Cravatt, B.F., 2004. Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, 43(45), 14332–14339.

Saito, N., Robert, M., Kitamura, S., Baran, R., Soga, T., Mori, H., Nishioka, T. & Tomita, M., 2006. Metabolomics approach for enzyme discovery. *J Proteome Res*, 5(8), 1979–1987.

Salama, N.R., Shepherd, B. & Falkow, S., 2004. Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *Journal of Bacteriology*, 186(23), 7926-35.

Sassetti, C.M., Boyd, D.H. & Rubin, E.J., 2003. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol*, 48(1), 77–84.

Schellenberger, J. & Palsson, B.Ø., 2009. Use of randomized sampling for analysis of metabolic networks. *The Journal of Biological Chemistry*, 284(9), 5457-61.

Scholle, M.D. & Gerdes, S., 2008. Whole-genome detection of conditionnally essential and dispensable genes in Escherichia coli via genetic footprinting. Dans A. L. Osterman & S. Y. Gerdes, éd. *Microbial Gene Essentiality: Protocols and bioinformatics*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 83-102.

Schuetz, R., Kuepfer, L. & Sauer, U., 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol*, 3, 119.

Schwarz, R., Liang, C., Kaleta, C., Kühnel, M., Hoffmann, E., Kuznetsov, S., Hecker, M., Griffiths, G., Schuster, S. & Dandekar, T., 2007. Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics*, 8, 313.

Scott, C.C., Makula, S.R. & Finnerty, W.R., 1976. Isolation and characterization of membranes from a hydrocarbon-oxidizing Acinetobacter sp. *J Bacteriol*, 127(1), 469–480.

Segrè, D., Vitkup, D. & Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23), 15112–15117.

Segrè, D., Zucker, J., Katz, J., Lin, X., D'haeseleer, P., Rindone, W.P., Kharchenko, P., Nguyen, D.H., Wright, M.A. & Church, G.M., 2003. From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS*, 7(3), 301–316.

Segura, D., Mahadevan, R., Juárez, K. & Lovley, D.R., 2008. Computational and Experimental Analysis of Redundancy in the Central Metabolism of Geobacter sulfurreducens. *PLoS Comput Biol*, 4(2), e36.

Senger, R.S. & Papoutsakis, E.T., 2008. Genome-scale model for Clostridium acetobutylicum: Part I. Metabolic network resolution and analysis. *Biotechnol Bioeng*, 101(5), 1036–1052.

Serres, M.H., Goswami, S. & Riley, M., 2004. GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Res*, 32(Database issue), D300–D302.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498–2504.

Shlomi, T., Berkman, O. & Ruppin, E., 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21), 7695–7700.

Simão, E., Remy, E., Thieffry, D. & Chaouiya, C., 2005. Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E.coli. *Bioinformatics (Oxford, England)*, 21 Suppl 2, ii190-6.

Smith, V., Botstein, D. & Brown, P.O., 1995. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 92(14), 6479-83.

Snitkin, E., Dudley, A., Janse, D., Wong, K., Church, G. & Segrè, D., 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome Biol*, 9(9), R140.

Song, J., Ko, K.S., Lee, J., Baek, J.Y., Oh, W.S., Yoon, H.S., Jeong, J. & Chun, J., 2005. Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. *Molecules and Cells*, 19(3), 365-74.

Steinmetz, L.M., Scharfe, C., Deutschbauer, A.M., Mokranjac, D., Herman, Z.S., Jones, T., Chu, A.M., Giaever, G., Prokisch, H., Oefner, P.J. & Davis, R.W., 2002. Systematic screen for human disease genes in yeast. *Nat Genet*, 31(4), 400–404.

Stelling, J., 2004. Mathematical models in microbial systems biology. *Curr Opin Microbiol*, 7(5), 513–518.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F.J. & Doyle, J., 2004. Robustness of cellular functions. *Cell*, 118(6), 675–685.

Stephanopoulos, G.N., Aristidou, A.A. & Nielsen, J., 1998. *Metabolic engineering. Principles and methodologies.*, San Diego, CA, USA: Academic Press, Elsevier Science.

Steuer, R., 2006. Review: on the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform*, 7(2), 151–158.

von Stockar, U. & Liu, J., 1999. Does microbial life always feed on negative entropy? Thermodynamic analysis of microbial growth. *Biochimica Et Biophysica Acta*, 1412(3), 191-211.

Strathern, J.N., Jones, E.W. & Broach, J. éd., 1982. *Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression*, Cold Spring Harbor Laboratory Press,U.S.A.

Suthers, P.F., Dasika, M.S., Kumar, V.S., Denisov, G., Glass, J.I. & Maranas, C.D., 2009. A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. *PLoS Computational Biology*, 5(2), e1000285.

Tagkopoulos, I., Liu, Y. & Tavazoie, S., 2008. Predictive behavior within microbial genetic networks. *Science*, 320(5881), 1313–1317.

Taylor, W.H. & Juni, E., 1961a. Pathways for biosynthesis of a bacterial capsular polysaccharide. I. Characterization of the organism and polysaccharide. *Journal of Bacteriology*, 81, 688-93.

Taylor, W.H. & Juni, E., 1961b. Pathways for biosynthesis of a bacterial capsular polysaccharide. II. Carbohydrate metabolism and terminal oxidation mechanisms of a capsuleproducing coccus. *Journal of Bacteriology*, 81, 694-703.

Taylor, W.H. & Juni, E., 1961c. Pathways for biosynthesis of a bacterial capsular polysaccharide. III. Syntheses from radioactive substrates. *The Journal of Biological Chemistry*, 236, 1231-4.

Thanassi, J.A., Hartman-Neumann, S.L., Dougherty, T.J., Dougherty, B.A. & Pucci, M.J., 2002. Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. *Nucleic Acids Research*, 30(14), 3152-62.

Thiele, I., Vo, T.D., Price, N.D. & Palsson, B.Ø., 2005. Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187(16), 5818–5830.

Thorne, K.J., Thornley, M.J. & Glauert, A.M., 1973. Chemical analysis of the outer membrane and other layers of the cell envelope of Acinetobacter sp. *J Bacteriol*, 116(1), 410–417.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Ménard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H. & Boone, C., 2004. Global mapping of the yeast genetic interaction network. *Science*, 303(5659), 808–813.

Tong, I.T., Liao, H.H. & Cameron, D.C., 1991. 1,3-Propanediol production by Escherichia coli expressing genes from the Klebsiella pneumoniae dha regulon. *Applied and Environmental Microbiology*, 57(12), 3541-6.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C., 2006. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, 34(1), 53–65.

Vallenet, D., Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., Bataille, E., Dossat, C., Gas, S., Kreimeyer, A., Lenoble, P., Oztas, S., Poulain, J., Segurens, B., Robert, C., Abergel, C., Claverie, J., Raoult, D., Médigue, C., Weissenbach, J. & Cruveiller, S., 2008. Comparative analysis of Acinetobacters: three genomes for three lifestyles. *PLoS ONE*, 3(3), e1805.

Vaneechoutte, M., Young, D.M., Ornston, L.N., De Baere, T., Nemec, A., Van Der Reijden, T., Carr, E., Tjernberg, I. & Dijkshoorn, L., 2006. Naturally transformable Acinetobacter sp. strain ADP1 belongs to the newly described species Acinetobacter baylyi. *Appl Environ Microbiol*, 72(1), 932–936.

Varma, A. & Palsson, B.O., 1994. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology*, 12, 994–998.

Villas-Boas, S.G., Roessner, U., Hansen, M.A.E., Smedsgaard, J. & Nielsen, J., 2007. *Metabolome Analysis: An Introduction*, Wiley InterScience.

Vojinović, V. & von Stockar, U., 2009. Influence of uncertainties in pH, pMg, activity coefficients, metabolite concentrations, and other factors on the analysis of the thermodynamic feasibility of metabolic pathways. *Biotechnology and Bioengineering*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19365870 [Accédé Avril 16, 2009].

de Vries, J. & Wackernagel, W., 2002. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), 2094-2099.

Vyazmensky, M., Sella, C., Barak, Z. & Chipman, D.M., 1996. Isolation and characterization of subunits of acetohydroxy acid synthase isozyme III and reconstitution of the holoenzyme. *Biochemistry*, 35(32), 10339–10346.

Whelan, K.E. & King, R.D., 2008. Using a logical model to predict the growth of yeast. *BMC Bioinformatics*, 9, 97.

Williams, P.A. & Ray, C.M., 2008. Catabolism of Aromatic Compounds by Acinetobacter. Dans U. Gerischer, éd. *Acinetobacter Molecular Biology*. Norfolk, UK: Caister Academic Press, pp. 99–117.

Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J. & Rojas, I., 2006. SABIO-RK: integration and curation of reaction kinetics data. *Lecture Notes in Computer Science*, 4075, 94.

Wunderlich, Z. & Mirny, L.A., 2006. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J*, 91(6), 2304–2311.

Yamazaki, Y., Niki, H. & Kato, J., 2008. Profiling of Escherichia coli Chromosome database. Dans A. L. Osterman & S. Y. Gerdes, éd. *Microbial Gene Essentiality: Protocols and bioinformatics*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 385-9.

Young, D.M., Parke, D. & Ornston, L.N., 2005. Opportunities for genetic investigation afforded by Acinetobacter baylyi, a nutritionally versatile bacterial species that is highly competent for natural transformation. *Annu Rev Microbiol*, 59, 519–551.

Zaslaver, A., Mayo, A.E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M.G. & Alon, U., 2004. Just-in-time transcription program in metabolic pathways. *Nat Genet*, 36(5), 486–491.

Zhang, R. & Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research*, 37(Database issue), D455-8.

ANNEXE

Nous plaçons en annexes deux articles supplémentaires auxquels nous avons contribués. Le premier présente l'interface web CycSim de prédiction des phénotypes de croissance de mutants pour plusieurs organismes, le second est une revue des approches de biologie systémique appliquées à l'exploration du métabolisme d'*A. baylyi*.

Le Fèvre, F., Smidtas, S., Combe, C., Durot, M., d'Alché-Buc, F. & Schachter, V., 2009. CycSim - an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics*, 25(15), 1987-1988.

de Berardinis, V., Durot, M., Weissenbach, J. & Salanoubat, M., 2009. Acinetobacter baylyi ADP1 as a model for metabolic system biology. *Curr Opin Microbiol*, 12(5), 568-576.

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 15 2009, pages 1987–1988
doi:10.1093/bioinformatics/btp268

Systems biology

CycSim—an online tool for exploring and experimenting with genome-scale metabolic models

F. Le Fèvre[1], S. Smidtas[1], C. Combe[1,2], M. Durot[1], Florence d'Alché-Buc[2] and V. Schachter[1,*]

[1]CEA, DSV, IG, Genoscope, UMR 8030, Evry, F-91057 and [2]IBISC, FRE 3190 CNRS, Université d'Evry, Evry, France

Received on October 7, 2008; revised on March 20, 2009; accepted on April 17, 2009

Advance Access publication May 6, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: *CycSim* is a web application dedicated to *in silico* experiments with genome-scale metabolic models coupled to the exploration of knowledge from BioCyc and KEGG. Specifically, CycSim supports the design of knockout experiments: simulation of growth phenotypes of single or multiple gene deletions mutants on specified media, comparison of these predictions with experimental phenotypes and direct visualization of both on metabolic maps. The web interface is designed for simplicity, putting constraint-based modelling techniques within easier reach of biologists. CycSim also functions as an online repository of genome-scale metabolic models.

Availability: http://www.genoscope.cns.fr/cycsim

Contact: cycsim@genoscope.cns.fr

1 INTRODUCTION

Constraint-based modelling (Price *et al.*, 2004) is a framework, simple and abstract enough to allow tractable modelling of metabolism at genome-scale, providing direct insights into the genotype–phenotype relationship. Constraint-based models (CBM) consist of a stoichiometric representation of the whole-cell metabolism together with a set of constraints on reaction fluxes. A wide variety of computational methods have been developed for this framework to characterize metabolic capabilities, help to discover new reactions, simulate scenarios of metabolic evolution or design experimental strategies to investigate metabolic behaviours (Feist and Palsson, 2008).

A few simulation tools (Becker *et al.*, 2007; Beste *et al.*, 2007; Lee *et al.,* 2003; Sympheny, www.genomatica.com) and model repositories (Le Novère *et al*., 2006; BiGG, unpublished data, http://bigg.ucsd.edu) have been proposed to the growing community of CBM users. These software tools have been limited in their usefulness to biologists for several reasons. First, most are either commercial, or add-ons to commercial platforms (e.g. MATLAB, http://www.mathworks.com). Next, they are typically directed at users with a background in modelling. Lastly, these tools are not designed to explore the biochemical and genomic knowledge underlying the metabolic models. Currently, the most convenient tools to reconstruct metabolic networks from genome annotation are databases of reference pathways such as BioCyc (Karp *et al.,* 2005) and KEGG (Kanehisa *et al.,* 2008). These databases provide descriptive and queriable views of the genetic and biochemical components of metabolism, but do not support modelling, simulation or prediction.

To address these shortcomings, we introduce *CycSim*, a web platform which supports *in silico* experiments with a variety of metabolic models, puts both the design and the results of these experiments in the visual context of reference pathways databases and allows confrontation with experimental data.

2 FUNCTIONALITIES

Predictions: CycSim supports *in silico* experiments with metabolic models. Each experiment consists in selecting a wild-type strain, choosing one or several genetic perturbations (e.g. knockout), and picking a set of growth media. Growth phenotype predictions are then generated for all (mutant, medium) pairs. These predictions can be compared against experimental growth phenotypes when available (Fig. 1). Two prediction methods are implemented: flux balance analysis and metabolites producibility check (Feist and Palsson, 2008). For any given (mutant, medium) pair, CycSim can also compute a flux distribution that is compatible with the model constraints and the objective function.

Visualisation: reactions, pathways and genes can be visualized in their context through a tight coupling of the CycSim core with the pathway display layers of BioCyc and KEGG. For instance, clicking on a reaction in the simulation panel will show the corresponding BioCyc reaction page augmented with information from the active model (i.e. balanced reaction equations or the Boolean gene-reaction correspondence). Conversely, a gene can be deleted from the current model by selecting it from a pathway map. Predictions and experimental results can be directly visualized and compared on pathways.

Model and data repository: the online CycSim repository stores information relative to three organisms: *Escherichia coli* (Feist *et al.,* 2007), *Saccharomyces cerevisae* (Duarte *et al.,* 2004) and *Acinetobacter baylyi ADP1* (Durot *et al.*, 2008). For each, CycSim includes (i) a genome-scale metabolic model; (ii) a detailed correspondence between that model and relevant data of that organism [EcoCyc, (Karp *et al.*, 2007); YeastCyc (Christie *et al.*, 2004); and AcinetoCyc (Durot *et al.*, 2008)]; (iii) a set of media definitions; and (iv) experimental growth phenotype datasets. Altogether, CycSim includes 2800 genes, 3700 reactions, 1400 metabolites, 190 media, 20 000 experimental phenotypes and 550

*To whom correspondence should be addressed.

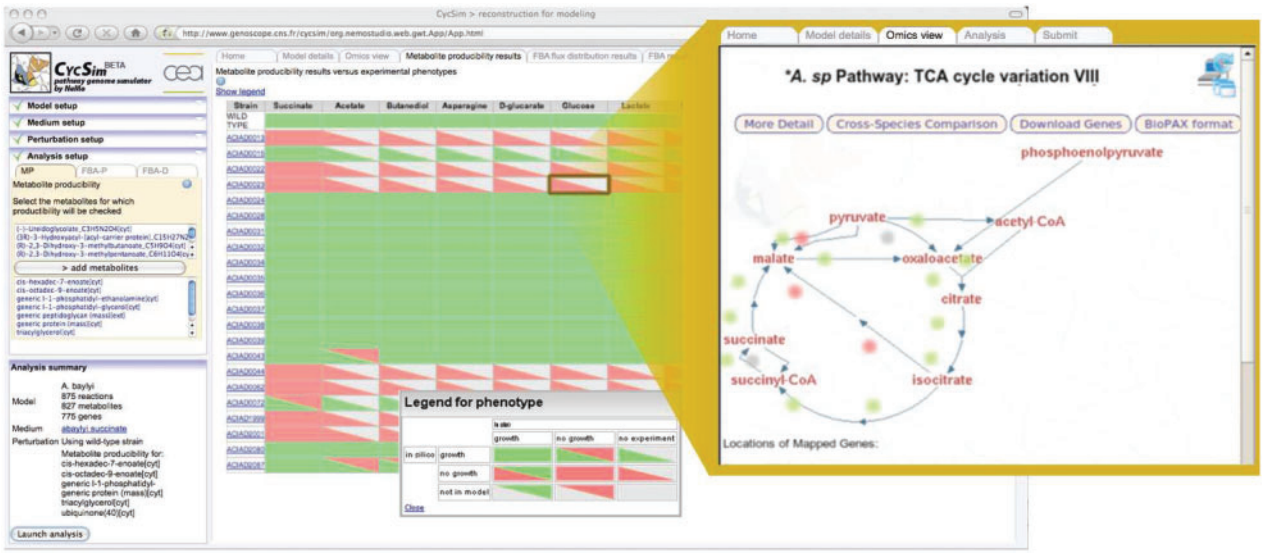


Fig. 1. CycSim screenshots. From the analysis of growth phenotypes of multiple mutants on multiple media (left), a flux distribution can be computed and visualized directly on relevant pathways (right).

pathways. Any of these four data types can be submitted online, using for models the SBML format, enhanced with MIRIAM annotations (Finney and Hucka, 2003; Le Novère *et al.,* 2005).

3 ARCHITECTURE AND TECHNOLOGIES

In order to facilitate operations from any computer, CycSim was developed as a web application using the AndroMDA framework (http://www.andromda.org) deployed on a Java application server (JBoss, http://www.jboss.org) with a MySQL backend (http://www.mysql.com). CycSim uses the AJAX technology (GWT, http://code.google.com/webtoolkit). In order to ensure the availability of sufficient computational resources, computations are performed on the server. A simple mechanism ensures some persistence of user sessions: the settings of each analysis are saved on the server and can be retrieved through a unique identifier.

In order to foster extensions by its developers or by the bioinformatics community, CycSim is based on a comprehensive UML model, which covers biochemical information (reactions and phenotype experiments) and information specific to CBM (fluxes and perturbations). Furthermore, web services are provided to programmatically access the models contained in CycSim (http://www.genoscope.cns.fr/cycsim/webservices.html).

4 CONCLUSIONS

CycSim is a simple online tool capable of handling several genome-scale metabolic models from a central repository in order to perform phenotype predictions, confronted to experimental data, and interpreted in the context of biological knowledge. CycSim facilitates the identification of inconsistencies, the design of new experiments and the iterative refinement of models using experimental data. We expect that the value of the biochemical insights obtained using CycSim will rise as more metabolic models are added to the repository, facilitating comparative analyses.

Funding: European FP6 Networks of Excellence BioSapiens (LSHG-CT-2003-503265); ENFIN (LSHG-CT-2005-518254).

Conflict of Interest: none declared.

REFERENCES

Becker,S. *et al.* (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protocols*, **2**, 727–738.

Beste,D.J. *et al.* (2007) GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism. *Genome Biol.*, **8**, R89.

Christie,K.R. *et al.* (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.

Duarte,N.C. *et al.* (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.*, **14**, 1298–1309.

Durot,M. *et al.* (2008) Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.*, **2**, 85.

Feist,A.M. and Palsson,B.Ø. (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.*, **26**, 659–667.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Finney,A. and Hucka,M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.*, **31**, 1472–1473.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.,* **36**, D480–D484.

Karp,P.D. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.,* **33**, 6083–6089.

Karp,P.D. *et al.* (2007) Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res.*, **35**, 7577–7590.

Le Novère,N. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.*, **23**, 1509–1515.

Le Novère,N. *et al.* (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, **34**, D689–D691.

Lee,S.Y. *et al.* (2003) MetaFluxNet, a program package for metabolic pathway construction and analysis, and its use in large-scale metabolic flux analysis of Escherichia coli. *Genome Inform.*, **14**, 23–33.

Price,N.D. *et al.* (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, **2**, 886–897.